Youth Smoking Survey 2004-2005

Microdata User Guide

Health Canada Santé Canada

www.yss.uwaterloo.ca

Table of Contents

1.0 Introduction

The Youth Smoking Survey is undertaken with the cooperation and support of Health Canada. The 2004-2005 Youth Smoking Survey (YSS) was implemented under the leadership of Dr. Steve Manske (Principal Investigator) with the Centre for Behavioural Research and Program Evaluation (CBRPE) at the University of Waterloo, and Drs. Steve Brown and Mary Thompson (Co-Principal Investigators) from the Statistics and Actuarial Sciences Department at the University of Waterloo. The 2004-2005 YSS was coordinated by staff from the Population Health Research Group (PHR). The investigators and staff were assisted by a consortium of university and non-governmental organizations across the country:

Dr. Shirley Solberg, Memorial University
Ms. Meg McCallum, Canadian Cancer Society, Nova Scotia Division
Ms. Donna Murnaghan, University of Prince Edward Island
Dr. William Morrison, University of New Brunswick
Dr. Jennifer O'Loughlin, Institut national de santé publique du Québec
Dr. Dexter Harvey, Canadian Cancer Society, Manitoba Division
Ms. June Blau, Saskatchewan Coalition for Tobacco Reduction
Dr. Cameron Wild, University of Alberta
Dr. Chris Lovato, University of British Columbia

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

Population Health Research Group
Rashid Ahmed
Senior Data Analyst
Room LHN 2704
200 University Ave. W
Waterloo, Ontario N2L 3G1
University of Waterloo
Telephone: (519) 885-1211 ext. 6632
Fax: (519) 746-8171
E-mail: r4ahmed@uwaterloo.ca

Health Canada
Dr. Alan Diener
Office of Research
Surveillance and Evaluation
Tobacco Control Programme
123 Slater Street,
Ottawa, Ontario K1A 0K9
Telephone: (613) 957-7852
Fax: (613) 954-2292
E-mail: Alan_Diener@hc-sc.gc.ca

This manual has been produced to facilitate the manipulation of the micro-data file of the survey results. Please become familiar with the contents of this document before publishing or otherwise releasing any estimates derived from the microdata file of the 2004-2005 Youth Smoking Survey.

2.0 Background

The Youth Smoking Survey (YSS) is a Health Canada sponsored survey of students in grades 5 through 9 and their parents. The YSS was first administered in 1994 and was the largest and most comprehensive survey on youth smoking behaviour since 1979. The YSS was repeated in 2002 in order to track changes in the attitudes and behaviour of Canadian children and adolescents with respect to tobacco. At that time, the YSS was planned as a biennial survey.

The YSS is a unique and important survey because reliable data on smoking prevalence rates among young Canadians are somewhat sparse; this is especially true for adolescents in grades 5 through 9 who are most vulnerable to start smoking. Some information on smoking prevalence is obtained routinely through school surveys; however, national level data are obtained only rarely. Monitoring surveys, which address a variety of psychoactive substances, provide only limited information on important psychosocial and environmental factors, and rarely address behaviours, knowledge and attitudes in a way that allows evaluation of the impact of prevention efforts.

The main objective of the YSS is to provide benchmark data on national prevalence rates for students in grades 5 through 9. In addition to prevalence rates, the YSS also offers a detailed snapshot of purchasing behaviour, and information about the effect of continued tobacco marketing. This information is critical to assessing the need for increased legislative controls on tobacco, and bolstering public support for these policy options. Without this type of monitoring, the effectiveness of our prevention efforts cannot be gauged.

The YSS also provides a unique opportunity to advance our knowledge of the psychosocial correlates of smoking behaviour including initiation and cessation, as well as individual differences in the influence of tobacco marketing and purchasing controls. Questions regarding alcohol and other drug use, permit similar reporting on a national basis specific to youth in grades 7 through 9 (secondary I to III in Québec).

Finally, the collection of data from parents at the same time as youth responses is unique in a national survey of this type. The parent data will help in the investigation of social influence on youth smoking behaviour, provide socio-economic status items, and describe potential for exposure to tobacco smoke in private homes and cars.

Participating schools received a customized feedback report. The feedback report focuses on smoking and related behaviours and is intended to be relevant and prescriptive for schools and communities.

3.0 Concepts and Definitions

See section 6.4 (Creation of Derived Variables) for more detailed information about the derived variables reflecting these definitions.

Currently smokes

Has smoked at least 100 cigarettes in his/her lifetime, and has smoked in the 30 days preceding the survey.

Currently smokes daily: Has smoked at least 100 cigarettes in his/her lifetime, and has smoked at least one cigarette per day for each of the 30 days preceding the survey.

Currently smokes occasionally: Has smoked at least 100 cigarettes in his/her lifetime, and has smoked at least one cigarette during the 30 days preceding the survey, but has not smoked every day.

Formerly smoked

Has smoked 100 or more cigarettes in his/her lifetime but has not smoked at all during the 30 days preceding the survey.

Formerly smoked daily: Has smoked 100 or more cigarettes in his/her lifetime but has not smoked at all during the 30 days preceding the survey, and has at some time smoked every day for seven days in a row.

Formerly smoked occasionally: Has smoked 100 or more cigarettes in his/her lifetime but has not smoked at all during the 30 days preceding the survey, and has never smoked every day for seven days in a row.

Non Smoker

Has smoked fewer than 100 cigarettes in his/her lifetime.

Experimental smoker (beginner): Has smoked between 1 and 99 cigarettes in his/her lifetime, and has smoked in the 30 days preceding the survey.

Past experimenter: Has smoked between 1 and 99 cigarettes in his/her lifetime, but has not smoked in the 30 days preceding the survey.

Puffer: Has smoked less than one whole cigarette in his/her lifetime, but has tried smoking.

Never tried: Has never tried smoking, not even just a puff.

The YSS collects data from students in grades 5 through 9 or approximately ages 9 to 14. Given that smoking rates, as they are typically characterized, are very low in this age group, users may want to consider focusing on indicators such as “ever-tried” and susceptibility (refer to section 6.4 for a derived variable for susceptibility).

4.0 Survey Methodology

The 2004-2005 Youth Smoking Survey (YSS) was administered to a sample of children in grades 5 through 9 (primary 5, 6 and secondary I to III in Québec) by sampling from all public and private schools in Canada.

4.1 Population Coverage

All 10 provinces were selected to participate in the study by Health Canada. Sampling frames for each province began with a list of all school boards (alternatively called school divisions and school districts) in each of the provinces. Each provincial sampling frame consisted of a range of information about each eligible board (including board enrollment, health region, and number of schools).

4.2 Sample Design

The sampling of schools for the 2004-2005 YSS was conducted in two stages. At stage 1, school boards were sampled within each province. From the selected school boards, schools were then sampled. All students in grades 5 through 9 (primary 5, 6 and secondary I to III in Québec) in the selected schools were eligible for the final sample.

Stage 1 – Selection of School Boards

In Newfoundland and Labrador, Nova Scotia and Prince Edward Island, the small number of boards meant that all boards were selected for the sample. For the remaining provinces, the procedure described below was followed.

The Canadian Community Health Survey (CCHS) sampled sufficient adults in each health region to allow for an estimate of the current smoking rate at the level of the health region. Within each province, each school board was assigned an estimate of the adult smoking rate from the Canadian Community Health Survey (CCHS) for the public health region that contained the school board. In the case that a school board was located in more than one health region, an average smoking rate was calculated based on the smoking rates for each of the relevant health regions.

Within provinces, some school boards that had few schools were combined with school boards with a similar adult smoking rate and treated as a single board for sampling. The school boards were then rank ordered based on their adult smoking rates and each board was assigned to one of two strata so that approximately half the total student enrollment in any province was assigned to each stratum. The boards with the higher smoking rates formed the “Upper Stratum” and boards with the lower smoking rates formed the “Lower Stratum”.

Within each stratum in each province, boards were randomly selected with probability proportional to the total enrollment in the board in the numbers described below. After a first set of boards was selected, a second set of substitute boards was selected in case any board refused to participate.

Stage 2 – Selection of Schools (not including private or independent schools)

Within each selected board, schools were stratified into two strata. The Senior stratum contained schools with students in the senior elementary or high school grades, specifically schools with grades 5-8, 5-9, 6-8, 6-9, 7, 7-8, 7-9, 8, and 9. The Junior stratum contained those schools with students in grades 5, 6, 5-6, 5-7, and 6-7. To accommodate the larger number of students in Senior stratum schools, there was an over selection of Junior stratum schools where possible (i.e., the board included adequate schools in the Junior stratum). A second set of schools was also chosen to act as substitute schools in case any of the originally selected schools refused to participate.

Selection of Private and Independent Schools

Within each province, except British Columbia, lists of private schools were obtained. A simple random sample of private schools was selected in each province from these lists. The number of schools originally selected was roughly proportional to the number of students enrolled in private schools in that province. Substitute private schools were selected as replacement schools in case any selected private school refused to participate. In British Columbia, private schools are integrated into public school districts and so were eligible for selection as described above.

Replacement of Boards and Schools

If a selected board refused to participate, it was replaced from the substitute list with the board from the same stratum whose adult smoking rate was closest to that of the original school board. In British Columbia, all original and substitute school boards were approached. The high refusal rate meant a third round of board selections was required.

If a school refused, it was replaced with the next substitute school in the same stratum. Similarly, if a private school refused, it was replaced by the next substitute private school.

Selection of Students

Within each selected school, all students in the eligible grades were eligible for the final sample. Consent for student participation was obtained from the parents of students who participated in the survey. Students without consent were not allowed to complete a questionnaire.

Parent Interviews

In randomly selected classes (two per school/per grade), parents were invited to participate in a brief telephone interview. If there were only one or two classes in an eligible grade, these classes were automatically selected to participate in the parent interview. In some provinces, the number of grade 9 classes included in the parent interview condition was increased in order to obtain a satisfactory number of grade 9 parent interviews. In Ontario, a maximum of 4 grade 9 classes per school were selected and in the provinces of Alberta, Québec and Prince Edward Island, a maximum of 3 grade 9 classes per school were selected.

4.3 Sample Size

The following tables show the number of school boards, schools and students sampled for each province. More detail regarding response rates can be found in section 7.1.

Table 1: Board Recruitment Outcomes

	Approached	Agreed	Refused	Response Rate (%)
Newfoundland and Labrador	4	4	-	100
Nova Scotia	5	5	-	100
Prince Edward Island	2	2	-	100
New Brunswick	4	4	-	100
Québec	11	10	1	91
Ontario	14	12	2	86
Manitoba	9	8	1	89
Saskatchewan	6	6	-	100
Alberta	12	8	4	67
British Columbia	19	9	10	47
Canada	**86**	**68**	**18**	**79**

Table 2: School Recruitment Outcomes

	Approached	Agreed	Refused	Response Rate(%)
Newfoundland and Labrador	24	24	-	100
Nova Scotia	31	24	7	77
Prince Edward Island	25	24	1	96
New Brunswick	28	20	8	71
Québec	60	36	24	60
Ontario	106	41	65	39
Manitoba	32	28	4	88
Saskatchewan	32	22	10	69
Alberta	60	30	30	50
British Columbia	115	32	83	27
Canada	**513**	**281**	**232**	**55**

Table 3: Sample Size by Grade

	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Total
Newfoundland and Labrador	544	601	519	431	423	2518
Nova Scotia	517	622	590	530	523	2782
Prince Edward Island	656	781	394	360	294	2485
New Brunswick	420	588	603	580	390	2581
Québec	851	915	731	743	404	3644
Ontario	784	817	834	800	915	4150
Manitoba	619	706	610	638	437	3010
Saskatchewan	379	378	397	330	562	2046
Alberta	606	604	620	464	331	2625
British Columbia	505	645	596	988	668	3402
Canada	**5881**	**6657**	**5894**	**5864**	**4947**	**29243**

5.0 Data Collection

Data collection was conducted from February to June 2005 with school board and school recruitment beginning in October 2004. Students were surveyed in their classrooms. Parent interviews were completed by telephone within 3 weeks following the school data collection.

5.1 Questionnaire Design

Several key considerations guided the design of the student questionnaire:

- Comparability - the basis of the questionnaire was the 2002 YSS with most items unchanged to allow for comparisons with the 2002 and 1994 data;
- Responsiveness - to meet the needs of users of the data, provincial collaborators were given an opportunity to contribute topics/items;
- Relevance - to ensure value-added for participating schools, items important to schools were added in order to enhance school-level feedback reports.

The 2004-2005 student questionnaire was adapted from the 2002 YSS . A Content Committee which included Health Canada representatives and all provincial collaborators was formed to review and update the questionnaire. The provincial collaborators provided input on areas of interest for their province. The questionnaire was finalized through a series of reviews and meetings.

In October 2004, a pilot test of the student questionnaire and the protocols for data collection was conducted. As part of the pilot, students were asked to complete the questionnaire according to normal protocols, as well as provide feedback on any questionnaire items they found difficult to answer or did not understand. Two focus groups were held with small groups of students in order to explore reactions to the survey in more depth. Feedback was also collected from the school contact and teachers. As a result of the pilot test, refinements were made to six questions (e.g., rewording, additional response options) and to explanatory text (e.g., additional definitions, repetition of assurances of confidentiality).

The student questionnaire was formatted to be machine readable, so that data could be scanned directly into a computer. For some items, this necessitated the creation of response options to minimize the number of open-ended responses. The questionnaire contained no skip patterns (all students responded to all questions) in order that all students (regardless of behaviour) would finish at about the same time. A separate version of the questionnaire for grades 5-6 classes did not contain alcohol and drug items.

The 2004-2005 parent interview was also based on the 2002 survey. A number of changes were made in consultation with Health Canada representatives, including the addition of questions on parent marijuana use, and on smoking in vehicles.

5.2 Data Collection Protocols

Prior to implementation, all protocols and materials were approved by the University of Waterloo Human Research Ethics Committee. Local institutional review boards affiliated with the institutions of consortium members also reviewed the project at the provincial level where applicable (i.e., where collaborators were affiliated with universities).

Each provincial collaborator hired a site coordinator to be responsible for school board and school recruitment, data collection preparation and implementation. Site coordinators attended a two-day training session, participated in additional web-based training sessions, received a comprehensive manual and had ready access to the Student Data Collection Coordinator for advice regarding day-to-day issues. Materials, databases and protocols were centrally developed to ensure consistency across provinces.

Provincial site coordinators were responsible for all board and school recruitment. Before board recruitment began, project information packages were mailed to all provincial Ministries of Education. In all provinces, school boards were approached prior to contact with schools. A standard recruitment package included an invitation letter, a project summary, sample questionnaires, sample consent letters and forms, and a template feedback report. In addition, formal application forms and procedures were adhered to, as required by individual boards.

Active parental consent was required for participation in the student survey. A parent information letter and consent form was sent home with students. Parent information letters provided details about the project, contact information for project staff and referral to the website for further details including copies of the questionnaires. If a child's class was selected to participate in the parent interview, additional explanation and consent fields were included in the package. Parents were given a minimum of two weeks to return consent forms. To improve consent form return rates, schools chose to resend consent materials, conduct phone follow-up to parents, and/or provide verbal or written reminders to students.

Site coordinators worked with a school contact to arrange data collection at each school. School contacts were asked to provide a list of classes for the eligible grades that included: teacher name, course name and/or the classroom number, grade, room number (*optional)*, and the number of students enrolled. This information was used to prepare consent materials and was entered, along with other school particulars (e.g., address, data collection date, etc.) into a database. Upon receipt of consent materials, student information was entered into this database and questionnaire IDs were assigned. Questionnaires were bundled by classroom and couriered to the school contact for distribution to classroom teachers 1 to 2 days prior to the data collection date.

On the data collection date, teachers administered the survey according to detailed instructions, during a designated class period. The survey took on average 30 to 40 minutes to complete. To protect confidentiality, teachers were asked not to circulate among the students.

Completed questionnaires were placed in a classroom envelope. A project staff member (site coordinator or data collector) was required to attend each school data collection. The staff member set up a station in front of the school office or another central location. The data collector was available to answer questions and receive classroom envelopes at the end of the data collection period. Within a few days of data collection, the site coordinators shipped the completed questionnaires, organized by school and classroom, to the coordinating centre, the Population Health Research Group at the University of Waterloo, for processing.

Parent interviews were conducted through a sub-contract with Crawford Canada. Trained interviewers conducted the brief interview, usually within 3 weeks of the student data collection.

Interviewers made up to 25 call-backs over a 3 to 4 week period and provided a 24 hour toll free number for those who wished to call at a time convenient to them.

6.0 Data Processing

The main output of the Youth Smoking Survey (YSS) is a microdata file. This chapter presents a brief summary of the processing steps involved in producing this file.

6.1 Data Capture

Student questionnaires were machine scanned using Optical Mark Read (OMR) technology. Several quality control measures were used to ensure the accuracy of the scan data. First, all questionnaires were visually scanned and marks that were too light or incomplete (e.g., check marks instead of filled in circles) were darkened to ensure that they would be recognized by the scanner. During this process, the perforated booklets were separated and oriented in preparation for the OMR scan. Secondly, standards were used to ensure that the calibration of the scanner remained constant. Finally, all bundles of questionnaires were scanned twice and discrepancies were investigated. Staff were trained to make decisions according to strict criteria. For example, they had to distinguish between true uncodeable responses, not to be corrected (e.g., where the respondent chose two answers) and those which were machine errors that were to be corrected (e.g., where the respondent erased one mark and chose another answer, but the scanner picked up the erased mark too). There were a total of 712 cases of discrepancy checked and appropriate corrections made during the scanning process. Logbooks and a quality control record were kept to track the number of corrections made and to monitor the progress of merging files to create a school-level file.

6.2 Editing and Imputation

The following standard codes are used in the microdata file:

Valid skip - 6, 96 and 996

Don't know - 7, 97 and 997

Refused - 8, 98 and 998

Not stated - 9, 99 and 999

The YSS 2004-2005 student dataset had 30 761 records. One questionnaire was scanned twice and the duplicate file was removed. Students who were not in grade 5, 6, 7, 8, or 9 were removed from the data set, resulting in 29,570 eligible records.[1] Based on incomplete records, particularly related to the determination of smoking status, 327 individuals were removed. The final number of records is 29,243. Québec grades secondary I, II, III were converted to grades 7, 8, 9 respectively. The variable Y_q1 was changed to the converted grade values as D_grade. The variable Y_q1 was left for users who want to do analyses separately for Québec.

[1] The original dataset included an additional 1,190 students sampled in Ontario in grades 10, 11, and 12 for a related project.

The following items required specific editing and/or imputation:

Question 1 (Y_q1)

During the process of determining consent form return rates and consent rates it was determined that several students had filled out a grade on the survey that was inconsistent with the grades represented in the school. If a student filled out a grade that did not match the relevant grades in the school, the variable was recoded to match the nearest relevant grade. If the value of Y_q1 was changed by this process the indicator variable IMP_q1 was set to 1.

> If Y_q1 (grade) was missing, uncodeable or improper for the province, then the student's grade was found from another source. The primary source was the student consent form. This form was filled out by the parent or the student and then signed by the parent. "Grade" was one of the fields. The secondary source was the grade that was associated with the student's class identification number. If multiple grades were listed in these fields then the first grade listed was used. If the value of Y_q1 was changed based on these sources then the variable imp_q1 was set to 1; if no change was made IMP_q1 was set to 0.

Question 3 (Y_q3)

If Y_q3 (gender) was either missing or uncodeable, then gender was found from another source. The primary source was the student consent form. This form was filled out by the parent or the student and then signed by the parent. "Gender" was one of the fields. The secondary source was the student's name. If the name was unclear for gender, then gender was left missing. If the value of Y_q3 was changed based on these sources then the variable IMP_q3 was set to 1; if no change was made IMP_q3 was set to 0.

Question 23 (Y_q23)

This question asked how many cigarettes had been smoked on each of the last 7 days. The range that was allowed for each day was 0 – 36. All responses between 37 and 99 have been set to 99 (not stated). Valid skips and not stated responses were set to 96 and 99 as in other variables.

6.3 Coding of Open-ended Options

There were eight partially open-ended items on the student questionnaire that included an *Other (specify)* option. These write-in answers were examined and recoded or retained as *Other*. The recoding was done into existing categories or created answer categories. For example, 8 students wrote "the kind my friend has" (or a very similar variation) in the *Other* space for Q_27 even though "My friends smoke the same brand" was a response option. In this instance, 8 responses were recoded into the applicable response option.

6.4 Creation of Derived Variables

A number of variables in the microdata file have been derived by combining items on the questionnaire in order to facilitate data analysis. Examples of derived variables include the average number of cigarettes smoked daily and the number of years the respondent smoked.

DVTY1ST

1 = Current Smoker
2 = Former Smoker
3 = Never Smoker

1 = Current Smoker
A current smoker is a person who reports having smoked 100 cigarettes and has smoked in the past 30 days.

Y_q20: *Have you ever smoked 100 or more whole cigarettes in your life?*
= 1 (Yes)

And

Y_q24: *On how many of the last 30 days did you smoke one or more cigarettes?*
= 2 (1 to 5 days) or 3 (6 to 10 days) or
4 (11 to 20 days) or 5 (21 to 29 days) or
6 (30 days *(every day)*)

2 = Former Smoker
A former smoker is a person who reports having smoked 100 or more cigarettes but did not smoke in the last 30 days.

Y_q20: *Have you ever smoked 100 or more whole cigarettes in your life?*
= 1 (Yes)

And

Y_q24: *On how many of the last 30 days did you smoke one or more cigarettes?*
= 1 (None)

3 = Never Smoker
A never smoker is a person who reports that they have not smoked 100 or more whole cigarettes in their life time but they might have smoked a whole cigarette.

Y_q20: *Have you ever smoked 100 or more whole cigarettes in your life?*
=2 (No)

Or

Have never smoked a whole cigarette

DVTY2ST

1 = Current Daily Smoker
2 = Current Occasional Smoker
3 = Former Daily Smoker
4 = Former Occasional Smoker
5 = Experimental Smoker (Beginner)
6 = Past Experimental Smoker
7 = Puffer
8 = Never Tried

1 = Current Daily Smoker

A current daily smoker is a person who reports currently smoking cigarettes every day.

Y_q20: *Have you ever smoked 100 or more whole cigarettes in your life?*
= 1 (Yes)

And

Y_q24: *On how many of the last 30 days did you smoke one or more cigarettes?*
= 6 (30 days *(every day)*)

2 = Current Occasional Smoker

A current occasional smoker is a person who currently smokes cigarettes but not every day.

Y_q20: *Have you ever smoked 100 or more whole cigarettes in your life?*
= 1 (Yes)

And

Y_q24: *On how many of the last 30 days did you smoke one or more cigarettes?*
= 2 (1 to 5 days) or 3 (6 to 10 days) or
4 (11 to 20 days) or 5 (21 to 29 days)

3 = Former Daily Smoker

A former daily smoker is a person who smoked at least 100 cigarettes in his/her life time and smoked at least seven days in a row but did not smoke in the last 30 days.

Y_q20: *Have you ever smoked 100 or more whole cigarettes in your life?*
= 1 (Yes)

And

Y_q24: *On how many of the last 30 days did you smoke one or more cigarettes?*
= 1 (None)

And

Y_q21: *Have you ever smoked everyday for at least 7 days in row?*
=1 (Yes)

4 = Former Occasional Smoker

A former occasional smoker is a person who smoked at least 100 cigarettes in his/her life time and did not smoke for at least seven days in a row and also did not smoke in the last 30 days.

Y_q20: *Have you ever smoked 100 or more whole cigarettes in your life?*
= 1 (Yes)

And

Y_q24: *On how many of the last 30 days did you smoke one or more cigarettes?*
= 1 (None)

And

Y_q21: *Have you ever smoked everyday for at least 7 days in row?*
=2 (No)

5 = Experimental Smoker (Beginner)
An experimental smoker is a person who has smoked in the last 30 days but has not smoked 100 or more cigarettes.

Y_q18: *Have you ever smoked a whole cigarette?*
= 1 (Yes)

And

Y_q20: *Have you ever smoked 100 or more whole cigarettes in your life?*
= 2 (No)

And

Y_q24: *On how many of the last 30 days did you smoke one or more cigarettes?*
= 2 (1 to 5 days) or 3 (6 to 10 days) or
4 (11 to 20 days) or 5 (21 to 29 days) or
6 (30 days *(every day)*)

6 = Past Experimental Smoker
A past experimental smoker is a person who has smoked a whole cigarette but did not smoke in the last 30 days and also did not smoke 100 cigarettes in his/her life time.

Y_q18: *Have you ever smoked a whole cigarette?*
= 1 (Yes)

And

Y_q20: *Have you ever smoked 100 or more whole cigarettes in your life?*
= 2 (No)

And

Y_q24: *On how many of the last 30 days did you smoke one or more cigarettes?*
= 1 (None)

7= Puffer
A puffer is a person who has tried smoking, but has never smoked a whole cigarette.

Y_q11: *Have you ever tried cigarette smoking, even just a few puffs?* = 1 (Yes)

And

Y_q18: *Have you ever smoked a whole cigarette?* = 2 (No)

8 = Never Tried
A person classified as never tried, has never tried a cigarette, not even just a few puffs.

Y_q11: *Have you ever tried cigarette smoking, even just a few puffs?* = 2 (No)

P_SMOKE

1 = At least 1 parent smokes
0 = Neither parent smokes

1 = At least 1 parent smokes
If Y_q49: *Does your father, or the person who is like your father, smoke cigarettes?*
= 4 (He smokes now)

OR

Y_q51: *Does your mother, or the person who is like your mother, smoke cigarettes?*
= 4 (She smokes now)

If both Y_q49 and Y_q51 are; "not stated" then P_smoke = 99 (Not Stated).

If both Y_q49 and Y_q51 are; "do not know" then P_smoke= 97 (Do not know).

All other combinations of responses have a value of =0 (Zero).

SIB_SMOKE

1 = At least 1 sibling smokes
0 = No siblings smoke

1 = At least 1 sibling smokes
If Y_q53: *Do any of your sisters smoke cigarettes?*
= 4 (At least 1 of my sisters smokes now)

Or

If Y_q54: *Do any of your brothers smoke cigarettes?*
= 4 (At least 1 of my brothers smokes now)

If both Y_q53 and Y_q54 are; "not stated" then Sib_smoke = 99 (Not Stated).

If both Y_q53 and Y_q54 are; "do not know" then P_smoke = 97 (Do not know).

All other combinations of responses have a value of = 0 (Zero).

D_SUSCEPTIBLE

1 = No
2 = Yes
99 = Not Stated

The susceptibility scale is based on the following three questions.

Y_q14	*Do you think in the future you <u>might try</u> smoking cigarettes?*	1 = I have already tried smoking 2 = Definitely yes 3 = Probably yes 4 = Probably not 5 = Definitely not 99 = Not Stated
Y_q15	*If one of your best friends was to offer you a cigarette would you smoke it?*	1 = Definitely yes 2 = Probably yes 3 = Probably not 4 = Definitely not 99 = Not Stated
Y_q16	*At anytime during the <u>next year</u> do you think you will smoke a cigarette?*	1 = Definitely yes 2 = Probably yes 3 = Probably not 4 = Definitely not 99 = Not Stated

- If Y_q14 = 5 and Y_q15 = 4 and Y_q16 = 4 then D_SUSCEPTIBLE = 1;
- If Y_q14 = 1,2,3 or 4 **And** Y_q15 = 1,2 or 3 **And** Y_q16 = 1,2 or 3 then D_SUSCEPTIBLE = 2;
- If Y_q14 = 99 **Or** Y_q15 = 99 **Or** Y_q16 = 99 then D_SUSCEPTIBLE = 99.

DVSELF

The objective of this variable is to measure the student's overall self-esteem and is based on the following item:

Y_q8: *Choose the answer that best describes how you feel.*
a) In general, I like the way I am
b) Overall, I have a lot to be proud of
c) A lot of things about me are good
d) When I do something, I do it well
e) I like the way I look

1 = False
2 = Mostly False
3 = Sometimes False / Sometimes True
4 = Mostly True
5 = True

The above scale for the 5 parts of Y_q8 was recoded as follows:
0 = False
1 = Mostly False
2 = Sometimes False / Sometimes True
3 = Mostly True
4 = True

Then the scores were added up and averaged across the questions that were answered by the student giving an overall score for variable DVSELF.

The next five derived variables use the following questions from the survey.

Think back over the last 7 days. Find yesterday on the wheel and fill in the number of cigarettes that you smoked. Then follow the wheel backwards and fill in the number of cigarettes you smoked on each of the last 7 days. a) Sunday b) Monday c) Tuesday d) Wednesday e) Thursday f) Friday g) Saturday *Coverage: Respondents where Y_q18=1 (Whole Cigarette)*	 Y_q23sun Y_q23mon Y_q23tue Y_q23wed Y_q23thu Y_q23fri Y_q23sat	 0 = 0 Cigarettes smoked 1 : 36 Cigarettes smoked 96 = Valid skip 99 = Not Stated

DVAMTSMK
The average number of cigarettes smoked per day in the past week as an integer value:

$$\frac{(Y\_q23sun + Y\_q23mon + Y\_q23tues + Y\_q23wed + Y\_q23thurs + Y\_q23fri + Y\_q23sat)}{7}$$

All responses had to have valid responses for valid data.

If all responses have 99 or if any of the days are missing then DVAMTSMK=99.

DVCIGWK
Total number of cigarettes smoked in the 7 days prior to the survey.

Y_q23sun + Y_q23mon + Y_q23tues + Y_q23wed + Y_q23thurs + Y_q23fri + Y_q23sat

Not necessary for all to have valid responses.

Zero value has been treated as a valid response.

If all days have missing data then DVCIGWK=99.

DVNDSMK
Number of days on which respondent smoked in the week prior to the survey.

A count of Y_q23sun, Y_q23mon, Y_q23tues, Y_q23wed, Y_q23thurs, Y_q23fri, and Y_q23sat excludes the missing responses.

Zero has been treated as a zero response.

DVAVCIGD
Average number of cigarettes smoked on the days that the respondent smoked.

$$\frac{\text{DVCIGWK}}{\text{DVNDSMK}}$$

DVSMKPTN

1 = Smoked every day
2 = Smoked week days only
3 = Smoked weekend days only
4 = Did not smoke in the last 7 days
5 = Other pattern
99 = Not stated

Calculated based on these variables:
Y_q23sun, Y_q23mon, Y_q23tues, Y_q23wed, Y_q23thurs, Y_q23fri , Y_q23sat

6.5 Skip Patterns

The youth questionnaire was intentionally designed with no respondent-use skip patterns to avoid the identification of smokers by rate of completion during the classroom session. Thus all smoking behaviour items included a response option such as, *I do not smoke.* However, due to the logical flow of the questions, a number of questions are extraneous based on the answer to a previous question. In these cases, a skip pattern has been imposed onto the data set. If a question could have been skipped, if this were allowable within the structure of the questionnaire, it was coded as 96 or 996. The following explains each question that has a 96 or a 996 code and the logical reasoning for coding the question in that way.

Y_q12 *(How old were you when you first tried smoking cigarettes, even just a few puffs?)* is only relevant if the respondent had tried smoking cigarettes. Therefore, Y_q12 was coded as 96 *(Valid Skip)* if Y_q11 *(Have you ever tried cigarette smoking, even just a few puffs?)* was 2 *(No)*.

Y_q18 *(Have you ever smoked a whole cigarette?)* is only relevant if the respondent had tried smoking cigarettes. Therefore, Y_q18 was coded as 96 *(Valid Skip)* if Y_q11 *(Have you ever tried cigarette smoking, even just a few puffs?)* was 2 *(No)*.

Y_q19 *(How old were you when you smoked your first whole cigarette?)* is only relevant if the respondent had smoked a whole cigarette. Therefore, Y_q19 was coded as 96 *(Valid Skip)* if Y_q18 *(Have you ever smoked a whole cigarette?)* was 2 *(No)* or 96 *(Valid Skip)*.

Y_q20 *(Have you ever smoked 100 or more whole cigarettes in your life?)* is only relevant if the respondent had smoked a whole cigarette. Therefore, Y_q20 was coded as 96 *(Valid Skip)* if Y_q18 *(Have you ever smoked a whole cigarette?)* was 2 *(No)* or 96 *(Valid Skip)*.

Y_q21 *(Have you ever smoked every day for at least 7 days in a row?)* is only relevant if the respondent had tried smoking cigarettes. Therefore, Y_q21 was coded as 96 *(Valid Skip)* if Y_q11 *(Have you ever tried cigarette smoking, even just a few puffs?)* was 2 *(No)*.

Y_q22 *(How old were you when you first smoked every day for at least 7 days in a row?)* is only relevant if the responded had smoked every day for at least 7 days. Therefore, Y_q22 was coded as 96 *(Valid Skip)* if Y_q21 *(Have you ever smoked every day for at least 7 days in a row?)* was 2 *(No)* or 96 *(Valid Skip)*.

Y_q23sun to Y_q23sat *(Think back over the last 7 days. Fill in the number of cigarettes you smoked on each of the last 7 days.)* are relevant only if the respondent had smoked a whole cigarette. Therefore, Y_q23sun to Y_q23sat were coded as 96 *(Valid Skip)* if Y_q18 *(Have you ever smoked a whole cigarette?)* was 2 *(No)* or 96 *(Valid Skip)*.

Y_q24 *(On how many of the last 30 days did you smoke one or more cigarettes?)* and Y_q25 *(Thinking back over the last 30 days, on the days that you smoked, how many cigarettes did you usually smoke each day?)* are only relevant if the respondent had smoked a whole cigarette. Therefore, Y_q24 and Y_25 were coded as 96 *(Valid Skip)* if Y_q18 *(Have you ever smoked a whole cigarette?)* was 2 *(No)* or 96 *(Valid Skip)*.

Y_q26 *(What brand of cigarettes do you usually smoke?)* is only relevant if the respondent had tried smoking cigarettes. Therefore, Y_q26 was coded as 96 *(Valid Skip)* if Y_q11 *(Have you ever tried cigarette smoking, even just a few puffs?)* was 2 *(No)*.

Y_q27a to Y_q27j *(Why do you smoke the brand of cigarettes that you do?)* and Y_q28 *(During the past 12 months, have you switched cigarette brands?)* are only relevant if the respondent reported that they have a usual brand. Therefore, Y_q27a to Y_q27j and Y_q28 were coded as 96 *(Valid Skip)* if Y_q26 was 1 *(I do not smoke)*, 2 *(I do not have a usual brand)*, 96 *(Valid Skip)* or 99 *(Not Stated)*.

Y_q29 *(Where do you usually get your cigarettes?)* and Y_q30a to Y_q30f *(How do you go about buying cigarettes from a store?)* are only relevant if the respondent had tried smoking cigarettes. Therefore, Y_q29 and Y_q30a to Y_q30f were coded as 96 *(Valid Skip)* if Y_q11 *(Have you ever tried cigarette smoking, even just a few puffs?)* was 2 *(No)*.

Y_q35ba to Y_q35bd *(Do you sometimes buy single cigarettes? If you do, where do you buy them?)* are only relevant if the respondent had purchased single cigarettes. Therefore, Y_q35ba to Y_q35bd were coded as 96 *(Valid Skip)* if Y_q35a was 1 *(I do not smoke)*, 2 *(I do not buy single cigarettes)*, or 99 *(Not Stated)*.

Y_q36 to Y_q40 (cessation items) are only relevant if the respondent had tried smoking cigarettes. Therefore, Y_q36 to Y_q40 were coded as 96 *(Valid Skip)* if Y_q11 *(Have you ever tried cigarette smoking, even just a few puffs?)* was 2 *(No)*.

Y_q41 to Y_q46 (addiction scale) are only relevant if the respondent had smoked in the last 30 days. Therefore, Y_q41 to Y_q46 were coded as 96 *(Valid Skip)* if Y_q24 was 1 (None) or 96 *(Valid Skip)*.

Y_q50 *(How does your father, or the person who is like your father, feel about your smoking?)* is only relevant if the respondent had tried smoking cigarettes. Therefore, Y_q50 was coded as 96 *(Valid Skip)* if Y_q11 *(Have you ever tried cigarette smoking, even just a few puffs?)* was 2 *(No)*.

Y_q52 *(How does your mother, or the person who is like your mother, feel about your smoking?)* is only relevant if the respondent had tried smoking cigarettes. Therefore, Y_q52 was coded as 96 *(Valid Skip)* if Y_q11 *(Have you ever tried cigarette smoking, even just a few puffs?)* was 2 *(No)*.

Y_q73 to Y_q88d_age (alcohol and drug use items) are only intended for respondents in grade 7,8 or 9. Therefore, Y_q73 to Y_q88d_age was coded as 96 *(Valid Skip)* if grade was indicated as 5 *(Grade 5)* or 6 *(Grade 6)*. Note that the module of the questionnaire distributed to grade 5 and 6 classes did not include these items.

Y_q74 *(How old were you when you first had a drink of alcohol that is more than a sip?)* is only relevant if the respondent had ever had a drink of alcohol. Therefore, Y_q74 was coded as 96 *(Valid Skip)* if Y_73 *(Have you ever had a drink of alcohol, that is more than just a sip?)* was 2 *(No)*, 96 *(Valid Skip)* or 99 *(Not Stated)*.

Y_q75 *(In the last year, how often did you drink alcohol?)* is only relevant if the respondent had ever had a drink of alcohol. Therefore, Y_q75 was coded as 96 *(Valid Skip)* if Y_73 *(Have you ever had a drink of alcohol, that is more than just a sip?)* was 2 *(No)*, 96 *(Valid Skip)* or 99 *(Not Stated)*.

Y_q76 *(Have you ever had 5 drinks or more of alcohol on one occasion?)* is only relevant if the respondent had ever had a drink of alcohol. Therefore, Y_q75 was coded as 96 *(Valid Skip)* if Y_73 *(Have you ever had a drink of alcohol, that is more than just a sip?)* was 2 *(No)*, 96 *(Valid Skip)* or 99 *(Not Stated)*.

Y_q77 *(How old were you when you first had 5 drinks or more of alcohol on one occasion?)* is only relevant if the respondent had ever had 5 drinks or more on one occasion. Therefore, Y_q77 was coded as 96 *(Valid Skip)* if Y_q76 *(Have you ever had 5 drinks or more of alcohol on one occasion?)* was 2 *(No)*, 96 *(Valid Skip)* or 99 *(Not Stated)*.

Y_q78 *(In the last year, how often did you have 5 drinks of alcohol or more on one occasion?)* is only relevant if the respondent had ever had 5 drinks or more on one occasion. Therefore, Y_q77 was coded as 96 *(Valid Skip)* if Y_q76 *(Have you ever had 5 drinks or more of alcohol on one occasion?)* was 2 *(No)*, 96 *(Valid Skip)* or 99 *(Not Stated)*.

Y_q81 *(How old were you when you first used marijuana or cannabis?)* is only relevant if the respondent had ever tried marijuana. Therefore, Y_q81 was coded as 96 *(Valid Skip)* if Y_q80 *(Have you ever used or tried marijuana or cannabis?)* was 2 *(No)*, 96 *(Valid Skip)* or 99 *(Not Stated)*.

Y_q82 *(In the last year, how often did you use marijuana or cannabis?)* is only relevant if the respondent had ever tried marijuana. Therefore, Y_q81 was coded as 96 *(Valid Skip)* if Y_q80 *(Have you ever used or tried marijuana or cannabis?)* was 2 *(No)*, 96 *(Valid Skip)* or 99 *(Not Stated)*.

DVAMTSMK *(The average number of cigarettes smoked per day in the past week.)* should only be determined if the respondent had smoked a whole cigarette. Therefore, DVAMTSMK was coded as 96 *(Valid Skip)* if Y_q18 *(Have you ever smoked a whole cigarette?)* was 2 *(No)* or 96 *(Valid Skip)*.

DVCIGWK *(Total number of cigarettes smoked in the past 7 days prior to the survey.)* should only be determined if the respondent had smoked a whole cigarette. Therefore, DVCIGWK was coded as 996 *(Valid Skip)* if Y_q18 *(Have you ever smoked a whole cigarette?)* was 2 *(No)* or 96 *(Valid Skip)*.

DVNDSMK *(Number of days on which respondent smoked in the week prior to the survey.)* should only be determined if the respondent had smoked a whole cigarette. Therefore, DVNDSMK was coded as 96 *(Valid Skip)* if Y_q18 *(Have you ever smoked a whole cigarette?)* was 2 *(No)* or 96 *(Valid Skip)*.

DVAVCIGD *(Average number of cigarettes smoked on the days that the respondent smoked.)* should only be determined if the respondent had smoked a whole cigarette. Therefore, DVAVCIGD was coded as 96 *(Valid Skip)* if Y_q18 *(Have you ever smoked a whole cigarette?)* was 2 *(No)* or 96 *(Valid Skip)*.

DVSMKPTN *(Smoking pattern in the last 7 days.)* should only be determined if the respondent had smoked a whole cigarette. Therefore, DVSMKPTN was coded as 96 *(Valid Skip)* if Y_q18 *(Have you ever smoked a whole cigarette?)* was 2 *(No)* or 96 *(Valid Skip)*.

6.6 Weighting

As described in section 4.2, the sampling of schools for the 2004-2005 YSS was conducted in two stages. At stage one, school boards were sampled within each province. From the selected school boards, schools were then sampled. All students in the grades 5 through 9 in the selected schools were eligible for the final sample.

The development of the survey weights was accomplished in stages. In the first stage a weight (W_1) was created based on the board selection scheme. In the second stage, a weight (W_2) was developed to account for the school selection. A third weight (W_3) was calculated to adjust for student non-response. At the second and third stages the survey weights were adjusted to remove undue variability and outlying weights that would have an undue influence on the resulting estimates and their estimated variances. The weights were also calibrated to the provincial gender and grade distribution so that the total of the survey weights by gender, grade and province would equal the actual enrollments in those groups. Finally, bootstrap weights were generated to attach to the data file.

Stage 1: Calculation of W_1

In Newfoundland, Nova Scotia and Prince Edward Island, all boards were selected and, hence, their first stage weight W_1=1. For the remaining provinces, all the boards in a given province were divided into two strata based on their sizes and related adult smoking rates (see section 4.2). Within each stratum, in each province, boards were randomly selected with probability proportional to the total enrollment in the board. For such a board *j*, W_1 has been computed as

$$\mathbf{W_1 = 1/\pi_j}$$

where π_j is the probability of inclusion for board *j* at stage 1, and where, except in British Columbia

$$\pi_j = \frac{lM_j}{\sum_{j=1}^{L} M_j};$$

in this expression, M_j = Total enrollment for board *j*

l = Number of boards selected in the stratum at the first stage of sampling, and

L = Total number of boards in the stratum

Because of a high number of refusals in BC, a second round of first stage sampling was necessary. This was accomplished by sampling additional boards to meet our minimum requirements. For these additional boards the probability of inclusion was calculated as $\pi_j^1 + \pi_j^2$, where

$$\pi^1_j = \frac{l_1 M_j}{\sum_{j=1}^{L} M_j}$$

$$\pi_j^2 = \sum_{s \not\ni j} p_1(s) * \frac{l_2 M_j}{\sum_{j \not\in s} M_{j'}}$$

and where

l_1 = Number of boards selected at the first round of first stage sampling
l_2 = Number of boards selected at the second round of first stage sampling
$p_1(s)$ = Probability of the sample *s* of school boards at the first round
π_j^1 = Probability of inclusion at the first round of first stage sampling

and

π_j^2 = Probability of inclusion at the second round of first stage sampling.

All private schools in a province were assigned to a distinct board. All the native school boards in Manitoba were assigned to one native school board. Because of very large enrollment, it was assumed that the Toronto District School Board would have been selected with certainty. Therefore their first stage weight W1 = 1.

The table below summarizes the number of school boards per stratum. As detailed in section 4.2, stratum 1 consists of school boards in public health regions with above average adult smoking rates as measured by the CCHS. Stratum 2 consists of school boards in public health regions with below average adult smoking rates. Stratum 3 includes private, French language and/or native school boards for provinces in which these are administratively separate from public boards. Stratum 3 also includes the three provinces (NL, NS, PE) in which all boards were selected.

Table 4: Number of boards selected by stratum:

	Stratum			Total
	1	**2**	**3**	
Newfoundland and Labrador	0	0	4	4
Nova Scotia	0	0	5	5
Prince Edward Island	0	0	2	2
New Brunswick	2	2	0	4
Québec	4	5	1	10
Ontario	5	6	1	12
Manitoba	3	3	2	8
Saskatchewan	2	3	1	6
Alberta	5	2	1	8
British Columbia	4	4	1	9
Canada	**25**	**25**	**18**	**68**

Stage 2: Calculation of W_2

Within each selected board, schools were stratified into two strata. Stratum 1, the Junior stratum, contained schools with student in grades 5, 6, 5-6, and 6-7. Stratum 2, the Senior stratum, contained schools with students in grades 5-8, 5-9, 6-8, 6-9, 7, 7-8, 7-9, 8, and 9. Schools were sampled by stratified random sampling without replacement. For the private schools, simple random sampling was used to select the required number of boards. The number of private schools selected in a province was proportional to the number of students enrolled in private schools in that province. Because the differences in sampling fractions in the two secondary strata within a board were administrative in origin, a raising factor from school to board was computed for each grade as follows:

$$W_2 = \frac{N(g)}{\sum_j [n_j(g)/c_j(g)]}$$

where

$n_j(g)$ is the number of students responding in grade g in school j

$c_j(g)$ is the consent rate for grade g in school j

N(g) is the enrollment in grade g in the board, obtained from administrative data

The table below summarizes the number of schools per stratum.

Table 5: Number of schools selected by stratum:

	Stratum		Total
	1 (Junior)	2 (Senior)	
Newfoundland and Labrador	12	12	24
Nova Scotia	11	13	24
Prince Edward Island	14	10	24
New Brunswick	8	12	20
Québec	24	12	36
Ontario	11	30	41
Manitoba	11	17	28
Saskatchewan	3	19	22
Alberta	10	20	30
British Columbia	16	16	32
Canada	**120**	**161**	**281**

Stage 3: Calculation of W_3

The adjustment for non-response was corrected by school at the grade (not class) level. The enrollment information collected by Site Coordinators was at the class level. However, some of the classes were split grades (e.g., grade 6 and grade 7 students were in the same class). The following method was used to determine the number of students enrolled in grades 5 through 9, the number of consent forms returned and the number of students who had parental consent to participate from the participating schools.

The enrollment, number of forms returned and positive consents were tabulated by using the grade level assigned to each class. A second table was generated to list all of the split class students to allow adjustments to be made if the actual grade of the student was different from the grade assigned to the student's class. There were several ways in which the actual grade of the student was determined. The primary method was to check the grade assigned on the student consent form. If this grade was an appropriate grade then this value was used as the student's actual grade. If the grade on the student's consent form was inappropriate or missing, then an attempt was made to link the student's identification number to the questionnaire data to see what the student filled out on the questionnaire. If it was not possible to link to the questionnaire data, then the actual grade was considered to be the grade assigned to the class.

Once the actual grade was determined, a table was generated to compare the actual grades to the grades assigned to the student's class. Adjustments were then made to the data by grade for the number of students enrolled, the number of returned consent forms and the number of positive consents, where appropriate.

During the process of determining consent form return rates and consent to participate rates, several students had filled out a grade on the survey that was inconsistent with the grades represented in the school. If a student filled out a grade that did not match the relevant grades in the school, it was recoded to match the nearest relevant grade; 17 records were so recoded.

W_3 is then calculated for each grade in each school as

$\mathbf{W}_3$ = (Number of eligible students with consent) number # of eligible students)/
(Number of eligible students)

However, among the W_3 computed this way there were some extreme weights due to low consent rates from certain schools. Using the rationale described by Potter (1988)[2], the lower level of response has been set to 0.25 for a given grade and school, and the weight components were recalculated.

The final un-calibrated weight is based on

$$\text{Weight} = W_1 * W_2 * W_3$$

[2] Potter, F. (1988). Survey of procedures to control extreme sampling weights. *ASA Proceedings of the Section on Survey Research Methods*, 453-458.

Stage 4: Calibration of survey weights

The weights just described are then calibrated using school administrative datasets that include the total student enrollment by gender and grade (grades 5 through 9) for each province. Province, grade and gender calibration are used to adjust the sampling weights so that estimated numbers of students in these domains reproduce known population numbers exactly.

Stage 5: Construction of Bootstrap Weights

The bootstrap weights for each province have been constructed separately as follows:

(i) Within each primary stratum (board stratum), the same number of boards is selected by simple random sampling (SRS) with replacement as has been selected in the original sample design.

(ii) Within each re-sampled board, and within each school stratum, the same number of schools is selected by simple random sampling (SRS) with replacement as has been selected in the original sample design.

(iii) Then within each re-sampled school, all eligible students who had consent to participate are selected.

(iv) The weights for re-selected units are recalculated and adjusted for the re-sampling inference based on the method of Rao and Wu (1988).[3]

(v) Finally the new weights are recalibrated to the provincial enrollment figures using the administrative datasets.

Six thousand (6000) bootstrap samples have been computed. The average of sets of twelve bootstrap weights have been used to create a set of 500 averaged bootstrap weights.

The formula for the weight adjustment is obtained as follows. Let w_{ijk} be the smoothed calibrated main weight for student *k* in school *j* of board *i.*

Let $\lambda_{1i} = \sqrt{\frac{n_i}{n_i - 1}(1 - \frac{n_i}{N_i})}$ where N_i is the number of boards in the primary stratum for board *i* and n_i is the number of boards actually selected in that stratum.

Let $\lambda_{2ij} = \sqrt{\frac{n_i}{N_i}(1 - \frac{m_{ij}}{M_{ij}})\frac{m_{ij}}{m_{ij} - 1}}$ where M_{ij} is the number of schools in the secondary stratum within board i for school *j,* and m_{ij} is the number of schools chosen in that secondary stratum within board i.

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

The bootstrap weight w^*_{ijk} would then be given by

$w_{ijk}[1 - \lambda_{1i} +$ (number of times board *i* has been resampled)* $(\lambda_{1i} - \lambda_{2ij})$ + (number of times school *j* has been resampled) * λ_{2ij}].

6.7 Suppression of Confidential Information

It should be noted that the Public Use Microdata Files (PUMF) may differ from the survey master files held by the Population Health Research Group. These differences usually are the result of actions taken to protect the anonymity of individual survey respondents. The most common actions are the suppression of file variables, grouping values into wider categories, and coding specific values into the "not stated" category. Specifically, the following variables have been removed from the PUMF: respondent's age, school board identifier, school identifier, class identifier, stratum identifier, postal code, and responses from the parent interviews.

7.0 Data Quality

There are various factors that influence data quality. This chapter summarizes threats to data quality and steps taken to ameliorate these.

7.1 Response Rates

There were various levels of non-response throughout the 2004-2005 YSS. A description of these levels is presented below along with the appropriate tables. First, some degree of non-response was noted among school boards and schools. Replacements were found for the majority of school boards and schools who refused to participate in the survey. The final response rates at the school board and school level, are presented in the tables below.

Table 6: Board Recruitment Outcomes

	Approached	Agreed	Refused	Response rate (%)
Newfoundland and Labrador	4	4	-	100
Nova Scotia	5	5	-	100
Prince Edward Island	2	2	-	100
New Brunswick	4	4	-	100
Québec	11	10	1	91
Ontario	14	12	2	86
Manitoba	9	8	1	89
Saskatchewan	6	6	-	100
Alberta	12	8	4	67
British Columbia	19	9	10	47
Canada	**86**	**68**	**18**	**79**

Table 7: School Recruitment Outcomes

	Approached	Agreed	Refused	Response rate (%)
Newfoundland and Labrador	24	24	-	100
Nova Scotia	31	24	7	77
Prince Edward Island	25	24	1	96
New Brunswick	28	20	8	71
Québec	60	36	24	60
Ontario	106	41	65	39
Manitoba	32	28	4	88
Saskatchewan	32	22	10	69
Alberta	60	30	30	50
British Columbia	115	32	83	27
Canada	**513**	**281**	**232**	**55**

The third level of response rate is based on individual student consent. The response rate at the student level is derived based on the number of eligible students as provided by school contacts for participating classes. Non-response at the student level can be attributed to several factors. Some parents/guardians refused to allow their child to take part in the survey, and even with parental consent some students refused to participate or the student was absent from class on the day of collection. Finally, some observations were removed because they did not contain sufficient information and could not be considered usable. (They were however, considered as valid responses in the calculation of the response rates.) The final response rates at the student level are summarized in the table below.

Table 8: Student Level Response Rates

	Eligible students	Students with consent	Completed questionnaires	Usable questionnaires	Response rate (%)*
Newfoundland and Labrador	4247	2773	2548	2518	60
Nova Scotia	5240	3067	2824	2782	54
Prince Edward Island	3637	2718	2509	2485	69
New Brunswick	4028	2937	2619	2581	65
Québec	7233	4022	3682	3644	51
Ontario	7386	4621	4188	4150	57
Manitoba	4739	3238	3034	3010	64
Saskatchewan	3280	2211	2070	2046	63
Alberta	4429	2874	2649	2625	60
British Columbia	7066	3701	3430	3402	49
Canada	**51285**	**32162**	**29553**	**29243**	**58**

***based on completed questionnaires.**

7.2 Survey Errors

The estimates derived from this survey are based on a sample of schools. Somewhat different estimates might have been obtained if a complete census had been taken using the same questionnaire, data collection staff, processing methods, and so on 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. Teachers 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 were taken 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 (a) the use of protocols that have been validated in previous studies of school-based data collection around youth smoking, (2) detailed instructions for teachers, (3) extensive training of project staff with respect to the survey procedures, (4) procedures to ensure that data capture errors were minimized, and (5) coding and edit quality checks to verify the processing logic.

8.0 Guidelines for Tabulation, Analysis and Release

Please note that this chapter is adapted from the 2002 Youth Smoking Survey User Guide written by Statistics Canada.[4] It details the guidelines to be adhered to by users tabulating, analyzing, publishing or otherwise releasing any data derived from the survey microdata files. With the aid of these guidelines, users of microdata should be able to produce the same figures as those produced by any statistician and, at the same time, will be able to develop currently unpublished figures in a manner consistent with these established guidelines.

8.1 Rounding Guide

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 un-rounded 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 un-rounded 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 un-rounded components and then are to be rounded themselves to the nearest 100 units (or the nearest one decimal) using normal rounding.

e) Under no circumstances are un-rounded estimates to be published or otherwise released by users. Un-rounded estimates imply greater precision than actually exists.

8.2 Sample Weighting Guidelines for Tabulation

The sample design used for the Youth Smoking Survey (YSS) 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

[4] Stats Canada (2002). Microdata User Guide: Youth Smoking Survey 2002. Accessible at: http://www.statcan.ca/english/sdds/document/4401_D2_T9_V2_E.pdf.

microdata files cannot be considered to be representative of the survey population, and will not correspond to those produced by Health Canada.

8.3 Definitions of Types of Estimates: Categorical and Quantitative

Before discussing how the YSS data can be tabulated and analyzed, 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 YSS.

8.3.1 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 students who ever smoked a whole cigarette or the proportion of smokers who usually buy single cigarettes from a friend or someone else 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: Where do you usually get your cigarettes?
R: I buy them from a vending machine/I buy them myself at a store / I buy them from someone /etc.

8.3.2 Quantitative Estimates

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

The only example of a quantitative estimate in the 2004-2005 YSS is the number of cigarettes smoked on each of the last seven days. If users want to estimate the average number of cigarette in a week, then the numerator is the total number of cigarettes smoked in the last seven days and the denominator would be the number of days smoked in last seven days.

8.3.3 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 $\hat{X}/\hat{Y}$ are obtained by:

a) summing the final weights of records having the characteristic of interest for the numerator ($\hat{X}$) ,

b) summing the final weights of records having the characteristic of interest for the denominator ($\hat{Y}$), then

c) dividing estimate a) by estimate b) ($\hat{X}/\hat{Y}$).

8.3.4 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 cigarettes smoked in the past seven days prior to the survey by students in grade 9 (secondary III in Québec) multiply the value reported in the derived variable DVCIGWK (number of cigarettes smoked in the past seven days prior to the survey) by the final weight for the record, then sum this value over all records with DVCIGWK < 996.

For example, to estimate the average number of cigarettes smoked in the past seven days prior to the survey by students in grade 9,

a) estimate the total number of cigarettes smoked in the past seven days prior to the survey by students in grade 9 ($\hat{X}$) as described above,

b) estimate the number of students in grade 9 (secondary III in Québec) ($\hat{Y}$) in this category by summing the final weights of all records with DVCIGWK < 996 then,

c) divide estimate a) by estimate b) ($\hat{X}/\hat{Y}$).

8.4 Guidelines for Statistical Analysis

The 2004-2005 YSS is based upon a complex sample design, with stratification, multiple stages of selection, and unequal probabilities of selection of respondents. Using data from such complex surveys challenges analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used. In order for survey estimates and analyses to be free from bias, the survey weights must be used.

While many analysis procedures found in statistical packages allow weights to be used, the meaning or definition of the weight in these procedures differ 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 less precise.

For example, suppose that analysis of all male respondents is required. The steps to rescale the weights are as follows:

1) select all respondents from the file who reported GENDER= male;

2) calculate the AVERAGE weight for these records by summing the original student weights from the microdata file for these records and then dividing by the number of respondents who reported GENDER = male;

3) for each of these respondents, calculate a RESCALED weight equal to the original student weight divided by the AVERAGE weight;

4) perform the analysis for these students using the RESCALED weight.

While this method produces reliable estimates of the coefficients under consideration in the analysis, note that because the stratification and clustering of the sample's design are still not taken into account, the variance estimates calculated in this way are likely to be under-estimates.

The calculation of more precise variance estimates requires detailed knowledge of the design of the survey. Such detail cannot be given in this microdata file to respect confidentiality. However, variances that take account for the sample design can be calculated from the bootstrap weights which are provided as a separate data file. Health Canada employed Stata for all analyses of the 2004-05 YSS. Variance estimates were conducted by using the BSWREG command. This procedure creates reliable estimates of the variance for both simple estimates such as totals, proportions and ratios and more complex analyses such as linear or logistic regression.[5] Another option is to use the Bootvar program available in both SAS and SPSS formats. It is made up of macros that compute variances for totals, differences between ratios and for linear and logistic regression. The bootstrap program can be requested free of cost from Stats Canada with the documents explaining how to modify and use the program to meet users' needs.

8.5 Coefficient of Variation Release Guidelines

Before releasing and/or publishing any estimate from the 2004-2005 YSS, users should first determine the quality level of the estimate. The quality levels are *acceptable*, *marginal* and *unacceptable*. Data quality is affected by both sampling and non-sampling errors as discussed in Chapter 7. However for this purpose, the quality level of an estimate will be determined only on the basis of sampling error as reflected by the coefficient of variation as shown in the table below. Nonetheless users should be sure to read Chapter 7 to be more fully aware of the quality characteristics of these data.

[5] BSWREG is a Stata ado file. For more information on the BSWREG command, including the necessary ado files to run the command please refer to the following two papers. Health Canada used the updated version of the BSWREG ado file that takes into account the fact that the bootstrap weights provided are actually mean weights. In the case of the 2004-05 YSS data, each of the 50 mean bootstrap weights is the average of twelve bootstrap weights. Hence when using the BSWREG command one must set cmeansb=12. Note that users must merge the YSS PUMF with the bootstrap weights data file before proceeding. This, as well as the other necessary details required to conduct this type of analysis, are provided in the following references:

Emmanuelle Piérard, Neil Buckley, and James Chowhan, Bootstrapping made easy: A STATA ADO file. *The Research Data Centres Information and Technical Bulletin.* Statistics Canada; Spring 2004, vol. 1 no.1, pp 20-36 (can be downloaded at http://www.statcan.ca/english/freepub/12-002-XIE/12-002-XIE2004001.pdf)

James Chowhan and Neil J. Buckley. Using mean bootstrap weights in Stata: A BSWREG revision. *The Research Data Centres Information and Technical Bulletin.* Statistics Canada; Spring 2005, vol. 2 no.1, pp 23-37. (can be downloaded at http://www.statcan.ca/english/freepub/12-002-XIE/12-002-XIE2005001.pdf)

First, the number of respondents who contribute to the calculation of the estimate should be determined. If this number is less than 30, the weighted estimate should be considered to be of unacceptable quality.

For weighted estimates based on sample sizes of 30 or more, users should determine the coefficient of variation of the estimate and follow the guidelines below. These quality level guidelines should be applied to weighted rounded estimates.

All estimates can be considered releasable. However, those of marginal or unacceptable quality level must be accompanied by a warning to caution subsequent users.

Table 9: Quality Level Guidelines

Quality Level of Estimate	Guidelines
1) Acceptable	Estimates have a sample size of 30 or more, and low coefficients of variation in the range of 0.0% to 16.5%. No warning is required.
2) Marginal	Estimates have a sample size of 30 or more, and high coefficients of variation in the range of 16.6% to 33.3%. Estimates should be flagged with the letter M (or some similar identifier). They should be accompanied by a warning to caution subsequent users about the high levels of error, associated with the estimates.
3) Unacceptable	Estimates have a sample size of less than 30, or very high coefficients of variation in excess of 33.3%. It is not recommended to release estimates of unacceptable quality. Such estimates should be replaced with the letter U (or some similar identifier) and the following statement: "Unreleaseable due to low sample size."

8.6 Release Cut-off's for the 2004-2005 Youth Smoking Survey Data

In the YSS 2002 Micro User's Guide, approximate sampling variability tables were supplied in order for the user to determine approximate Coefficients of Variation (C.V.'s) for estimates of different types. For the 2004-2005 Youth Smoking Survey data, users are encouraged to use bootstrap weights, as described earlier to calculate the variance of all estimates.

Coefficients of variation are derived using the variance formula for simple random sampling and incorporating a factor that reflects the underlying characteristics of the sampling design. The table below provides rough release cut-off's for population estimates for each province, based on coefficients of variation for estimates of population totals. For example, the table shows that the quality of a weighted estimate of 1,206 people possessing a given characteristic in Newfoundland and Labrador should be flagged as marginal.

Table 10: Release Cut-Off's by Province

Province	Acceptable		Marginal			Unacceptable	
	CV 0.0% – 16.5%		CV 16.6% – 33.3%			CV > 33.3%	
Newfoundland and Labrador	4,442	& over	1,206	to <	4,442	under	1,206
Prince Edward Island	538	& over	136	to <	538	under	136
Nova Scotia	1,333	& over	329	to <	1,333	under	329
New Brunswick	4,939	& over	1,301	to <	4,939	under	1,301
Quebec	43,586	& over	11,342	to <	43,586	under	11,342
Ontario	38,286	& over	9,647	to <	38,286	under	9,647
Manitoba	16,043	& over	4,676	to <	16,043	under	4,676
Saskatchewan	15,046	& over	4,409	to <	15,046	under	4,409
Alberta	8,886	& over	2,222	to <	8,886	under	2,222
British Columbia	19,879	& over	5,156	to <	19,879	under	5,156
Canada	**29,995**	**& over**	**7,360**	**to <**	**29,995**	**under**	**7,360**