



Microdata User Guide

Follow-up of Graduates Survey

Class of 2000

Public Use Microdata File



Statistics
Canada

Statistique
Canada

Canada

Table of Contents

1.0	Introduction.....	5
2.0	Background.....	7
3.0	Objectives.....	9
4.0	Content.....	11
4.1	Concepts and Definitions.....	11
4.2	Uses.....	13
5.0	Survey Methodology.....	15
5.1	Target Population.....	15
5.2	Survey Frame.....	16
5.3	Survey Design.....	16
5.3.1	Longitudinal Sample.....	16
5.3.2	Stratification.....	16
5.4	Sample Allocation, Selection and Size.....	17
6.0	Data Collection.....	21
6.1	National Graduates Survey (Class of 2000).....	21
6.2	Follow-up of Graduates Survey (Class of 2000).....	21
7.0	Data Processing.....	23
7.1	National Graduates Survey/Follow-up of Graduates (Class of 2000).....	23
7.1.1	Data Capture.....	23
7.1.2	Editing.....	23
7.1.3	Coding of Open-ended Questions.....	23
7.1.3.1	Coding of Education Programs.....	23
7.1.3.2	Coding of Industry and Occupation.....	24
7.1.3.3	Coding of “Other – Specify” Answers.....	24
7.1.4	Imputation.....	24
7.1.5	Creation of Derived Variables.....	25
8.0	Response Rates.....	27
9.0	Weighting and Sampling Variability.....	29
10.0	Data Quality.....	33
10.1	Sampling Errors.....	33
10.2	Non-sampling Errors.....	33
10.3	Non-response.....	34
10.4	Coverage.....	34
11.0	Guidelines for Tabulation Analysis and Release.....	37
11.1	Rounding Guidelines.....	37
11.2	Sample Weighting Guidelines for Tabulation.....	37
11.3	Definitions of Types of Estimates: Categorical and Quantitative.....	38
11.3.1	Tabulation of Categorical Estimates.....	39
11.3.2	Tabulation of Quantitative Estimates.....	39
11.4	Guidelines for Statistical Analysis.....	40
11.5	Release Guidelines.....	40
11.6	Release Cut-off's for the PUMF.....	41
12.0	Approximate Sampling Variability Tables.....	43
12.1	How to Use the Coefficient of Variation Tables for Categorical Estimates.....	43
12.1.1	Examples of Using the Coefficient of Variation Tables for Categorical Estimates.....	44
12.2	How to Use the Coefficient of Variation Tables to Obtain Confidence Limits.....	48
12.2.1	Example of Using the Coefficient of Variation Tables to Obtain Confidence Limits.....	49
12.3	How to Use the Coefficient of Variation Tables to Do a T-test.....	49
12.3.1	Example of Using the Coefficient of Variation Tables to Do a T-test.....	49
12.4	Coefficients of Variation for Quantitative Estimates.....	50
12.5	Coefficient of Variation Tables.....	50
13.0	Questionnaire, Code Sheets and Description of Derived Variables.....	51
14.0	Record Layout with Univariate Frequencies.....	53

1.0 Introduction

The Follow-up of Graduates Survey – Class of 2000 (FOG2000) was conducted by Statistics Canada from April 27th to July 24th, 2005. This manual has been produced to facilitate the manipulation of the public-use microdata file.

The public-use microdata file, or PUMF, contains a reduced list of variables compared to the FOG master file. The need to preserve the confidentiality of respondents dictated that many variables that could have been used to identify individuals (including all geographic information) be removed from the file. In addition, all continuous variables such as those relating to income, student loans or age at graduation, were converted to categorical variables, and many existing categorical variables were grouped into a smaller number of categories. Finally, local suppression was used where necessary to further protect confidentiality. Every effort was made to preserve the analytical utility of the data during this process.

It is also important to note that this PUMF contains fewer records than the master file. As an initial measure of diminishing the risk of disclosure, a subsample of the records from the master file was drawn. The PUMF therefore is made up of 11,200 records, or roughly half the number in the FOG master. Users should be aware that estimates produced using the subsample may not correspond exactly to those produced by Statistics Canada using the master file.

This document retains most of the content from the original user guide for the NGS and FOG master microdata file for informational purposes. Notes have been added to indicate where full content is not applicable to the PUMF.

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

Statistics Canada

Client Services
Centre for Education Statistics
Room SC-2000 B, Main Building
150 Tunney's Pasture Driveway
Ottawa, Ontario
K1A 0T6

Telephone: (613) 951-7608 or call toll-free 1 800 307-3382

Fax: (613) 951-4441

E-mail: educationstats@statcan.gc.ca

2.0 Background

In 1978, Statistics Canada conducted a survey on the labour market experiences of 1976 graduates from universities and community colleges in Canada. In 1984, a similar survey, the National Graduates Survey (NGS) of 1982 graduates, was sponsored jointly by the Department of the Secretary of State and Employment and Immigration Canada. The 1984 NGS expanded on the content of the previous survey and extended the population base to include those who completed trade/vocational programs in addition to community college and university graduates.

Since these two surveys in 1978 and 1984, a series of graduate surveys has been completed on the labour market experiences of university and community college graduates in Canada. The following is a summary of the graduate surveys conducted by Statistics Canada.

Graduation Year	Survey Year	Survey Name
1976	1978	Survey of 1976 Graduates of Post-Secondary Programs
1982	1984	Survey of 1982 Graduates (S82G) (also known as the National Graduates Survey or NGS)
1982	1987	Follow-up of 1982 Graduates (F82G)
1986	1988	Survey of 1986 Graduates (S86G)
1986	1991	Follow-up of 1986 Graduates (F86G)
1990	1992	Survey of 1990 Graduates (S90G)
1990	1995	Follow-up of 1990 Graduates (F90G)
1995	1997	Survey of 1995 Graduates (S95G)
1995	2000	Follow-up of 1995 Graduates (F95G)
2000	2002	National Graduates Survey - Class of 2000 (NGS2000)
2000	2005	Follow-up of Graduates Survey – Class of 2000 (FOG2000)

The Follow-up of Graduates Survey (Class of 2000), conducted from April 27th to July 24th 2005, updated the information obtained in the 2002 survey, covering the period between June 2002 and June 2005. The main content of the survey contains data on: the link between education experience and outcomes; information on the job held in the week prior to the interview; financial and loan information; additional education and training after graduation; and socio-economic background.

For the class of 2000, the content of the Survey of Graduates who Moved to the United States (conducted in 1999) was added to the traditional content of the NGS questionnaire so that graduates residing in the United States of America at the time of the 2002 survey were also interviewed. These graduates were subsequently interviewed at the time of the Follow-up of Graduates Survey (Class of 2000), as were graduates who had resided in Canada at the time of the 2002 survey, but who were residing in the United States of America at the time of the follow-up survey.

Note, however, that for confidentiality reasons, information specific to graduates who lived in the United States is not available on the PUMF.

Graduates from trade/vocational programs were interviewed in 2002 but not in the follow-up survey.

3.0 Objectives

The survey's primary objective is to obtain information on the labour market experiences of graduates entering the labour market, focusing on employment, occupations and the relationship between jobs and education.

The survey's key data objectives are:

- To obtain information for labour market analysis of a key youth group at an important time, focusing on education, training, employment, occupations and geographic mobility. The data and analysis will be useful for policy development.
- To obtain information on the exposure of graduates to additional learning opportunities.
- To extend available information required to improve occupational supply and demand projection models for various occupational categories.
- To obtain data regarding longer-term labour market experiences of graduates, with special emphasis on employment and occupations, for use in counseling on career and post-secondary education course selection.
- To obtain information on labour market experiences of members of target groups (such as women, native people and the disabled), which permits longitudinal and comparative analysis useful in the formulation of job equity policies.
- To gain a better understanding of school-work transitions and returns to human capital.
- To gain a better understanding of post-secondary education financing.
- To obtain more detailed information on knowledge and skills.

4.0 Content

The Follow-up of Graduates Survey – Class of 2000 (FOG2000) questionnaire is made up of 6 sections of questions. The following table describes the content of each section.

Section	Content
Graduates who live/lived in the United States (FUS) *	Graduates confirm whether or not they lived in the United States since the 2002 interview.
Activities last week (FLF)	Asks about the graduate's labour force activity the week before the 2005 interview.
Graduates who live/lived in the United States: activities prior to leaving and upon entry (FMU) *	Obtains information on activities of graduates in the United States since the 2002 interview as well as about their return to Canada, if applicable.
Education programs (FED) Education program description (FEDP)	Asks about educational programs taken and completed after the 2002 interview.
Student loans (FSL)	Asks questions about student loans and finances.
Demographics (FDE) Demographics roster (FDEM)	Asks general questions such as marital status, number of dependant children, income and disabilities.

*Note: this information is not available on the PUMF.

4.1 Concepts and Definitions

Graduation date

For the purpose of this survey, the graduation date is the year and month in which the graduate completed the requirements of his/her program. To complete the requirements of the program, graduates must have written and passed the last exam, submitted the last paper, report or project for a program, or defended a thesis. The variables PR_D11Y and PR_D11M from the National Graduates Survey contain the graduation date.

Graduates who moved to live in the United States of America

Graduates who live in the USA, or lived in the USA since the 2002 interview but have returned to Canada, are included in the survey. They may have moved to attend school, to work, or to accompany a partner or spouse. Anyone who visited or vacationed in the USA temporarily is not considered to have moved.

Transition after completing post-secondary studies

A number of modules in the survey are devoted to obtaining information on the graduate's activities after the 2002 interview. The information found in these modules allows for a detailed analysis on the graduate's transition after completing his/her post-secondary studies.

- The FLF module asks about the graduate's labour force activities during the week prior to the interview (i.e., employed, unemployed, or not in the labour force). Detailed information on the job held in the week prior to the interview is also collected.

- The FED and FEP modules collect information on educational programs taken and completed after the 2002 interview when these programs lead towards a diploma, certificate or degree that would take someone three months or more to complete if taken full-time.

Income

The income information is for the income received from all sources by the graduate in the calendar year 2004. It is not limited to monies that are taxable.

It includes:

- income from wages and salaries;
- net income from self-employment;
- regular Employment Insurance benefits as well as those for sickness, maternity or paternity leave, adoption, job creation, work sharing, retraining and benefits to self-employed fishermen;
- retraining and retirement benefits received under the Human Resources Development Canada employment insurance program;
- payments from provincial or municipal programs for persons in need such as Social Assistance or welfare;
- spousal support or child support;
- scholarships, grants, bursaries or fellowships;
- money from the Canada or Quebec Pension Plan;
- Canada Child Tax Benefits or provincial child tax benefits or credits;
- interest from Canadian and foreign sources;
- foreign dividends;
- taxable dividends received from Canadian corporations;
- net rental income;
- rents for leased farm land;
- regular income from an estate or trust fund;
- cash dividends from life insurance policies;
- pensions from deferred profit sharing plans and other private pension plans;
- money from parents, guardians or others that does not have to be repaid.

It excludes:

- monies received from student loans or any other loan;
- income tax refunds;
- tax-free Registered Retirement Savings Plan withdrawals used for purchasing a home;
- proceeds from the sale of property, businesses, financial assets or personal belongings;

- loans repaid to the graduate as a lender;
- refund of contributions to work-related pension plans.

4.2 Uses

Following from previous surveys, this survey extends the existing base of information on the labour-market experiences of recent graduates. Information derived from the survey has the potential to shed light on many areas of current interest. The following are examples of uses to which the survey's data is applied.

- The survey data can be used to update the occupational supply and demand models and the student flow model. These models project supplies of labour by occupation and industry, especially in highly-skilled and highly-qualified categories.
- Job equity programs will receive important labour market related information on designated groups such as women, aboriginal peoples, persons with disabilities and visible minorities.
- The survey provides concrete information regarding graduates' labour market experiences and career development during the five years after graduation. This information can be used to aid post-secondary education course selection and career counselling.

5.0 Survey Methodology

The Follow-up of Graduates Survey – Class of 2000 (FOG2000) is a longitudinal survey designed to collect data from Canadian graduates.

5.1 Target Population

The target population of the FOG2000 consists of all graduates from recognized public post-secondary Canadian colleges and universities having completed the requirements of an admissible program or obtained a diploma some time in 2000.

These graduates include:

- graduates of university programs that lead to bachelor's, master's or doctoral degrees, or that lead to specialised certificates or diplomas; and,
- graduates of post-secondary programs (that is, programs of one year or more in duration that normally require a secondary school completion or its equivalent for admission) in Colleges of Applied Arts and Technology (CAAT), Collèges d'enseignement général et professionnel (CEGEP in Quebec), community colleges, technical schools or similar institutions.

The survey excludes:

- graduates of skilled trades (that is, pre-employment programs that are normally three months or more in duration). A trade/vocational school is a public educational institution that offers courses to prepare people for employment in a specific occupation such as heavy equipment operator, automotive mechanic or upholsterer. Many community colleges and technical institutes offer certificates or diplomas at the trade level;
- graduates from private post-secondary institutions (for example, computer training and commercial secretarial schools);
- graduates who completed "continuing education" courses at universities and colleges (unless they led to a degree or diploma);
- graduates who took part-time trade courses (for example, adult education evening courses) while employed full-time;
- graduates who completed vocational programs that lasted less than three months or that were not in the skilled trades (for example, basic training and skill development); and
- graduates of apprenticeship programs.

5.2 Survey Frame

The survey frame for the 2000 graduates was created by Statistics Canada's Centre for Education Statistics from a list of all graduates from universities, colleges and trade/vocational schools in Canada.

Data on graduates were provided through two sources: the main source of information was from the individual institutions and provincial co-ordinating bodies, while the second source of graduate data came from the Postsecondary Student Information System (PSIS), which is maintained by the Centre for Education Statistics.

Where the PSIS data could not be extracted, files of graduates, preferably in electronic format, were requested from the institutions or provincial co-ordinating bodies. In a few cases, files were supplied by provincial Ministries (such as Ministère de l'éducation du Québec (MEQ), which provided data on its trade/vocational institutions, colleges and Université du Québec satellites) for some of the institutions in a province. The same information that is submitted to the PSIS was requested for each graduate: his/her name, permanent address and telephone number, local address and telephone number, qualification obtained in 2000, major field of study, date of birth, student number, immigration status, gender, mother tongue, graduation date and whether the program taken was a co-op program.

5.3 Survey Design

The NGS2000 and FOG2000 use a stratified simple random sample design without replacement of graduates within strata. The random selection was completed using a systematic method.

5.3.1 Longitudinal Sample

The survey involves a longitudinal design with graduates being interviewed at two different times: at two years and five years after graduating from post-secondary institutions in Canada. The sample design has been developed using a "funnel-shaped" approach, where only graduates that respond to the initial interview will be traced for the follow-up interview.

5.3.2 Stratification

There are three variables used for stratification; geographical location of the institution, level of certification and field of study. There are 13 geographical locations: the 10 provinces and the three northern territories. There are five levels of certification: trade/vocational programs, college programs, bachelor's degree, master's degree, and doctorate. As for the stratification level for the fields of study, it depends on the levels of certification. There are eight categories of field of study for the trade/vocational level and nine categories each for the college level and the three university level degrees (i.e., bachelor's, master's and doctorate) combined. Details about the field of study can be found in Appendix F. As with previous National Graduates Surveys (NGS), the field of study was obtained by grouping the Community College Student Information System (CCSIS) and the University Student Information System (USIS).

The combination of these variables makes for a total of 572 strata.

5.4 Sample Allocation, Selection and Size

The FOG2000 sample is a sub-sample of the NGS2000 sample, i.e comprised of NGS2000 respondents. Initially, the NGS2000 sample was divided into two components – the basic sample and the supplementary sample. The core sample was designed to yield estimates of a minimal proportion of 5.5% with a maximum coefficient of variation (CV) of 16.5% for any of the NGS2000's marginal. A marginal was defined as i) a given field of study regardless of the province of institution; or ii) a given province of institution regardless of the field of study; and that for each of the five levels of certification. The marginal's CVs were then allocated to each stratum (or cell in a table) to obtain the cells or strata CV using a raking-ratio algorithm. The last step consisted of converting the CV's into sample sizes.

The supplementary sample targeted specific sub-populations in order to meet the interests of external partners. The provinces of Quebec and Manitoba made such requests for graduates at the bachelor's and master's levels.

Finally, the last step consisted of over-sampling to compensate for expected non-response. The determination of the final sample size was based on some hypothesis about attrition rates for the follow-up survey and past NGS2000 response rates.

For FOG2000, it was determined that due to conceptual and sample requirement issues, it would be beneficial for the aims of the project as a whole to not follow-up with the Trade/vocational graduates who responded to the NGS2000. Moreover, as part of the survey, the respondent was asked to confirm the certification level. Therefore, the FOG2000 sample is comprised of all NGS2000 respondents whose reported variable indicated that they earned either a college diploma or certificate, a Bachelor's degree, a Master's degree or a Doctorate in 2000.

The table below presents the distribution of the population and the NGS sample size for the stratification variables.

The table in Section 8 provide the sample size and the number of responses by province and reported level of qualification for the Follow-up of 2000 Graduates. Detailed notes on the sample size and number of responses are provided at the beginning of the section.

Population and Sample Size by Province / Territory and Level of Certification *

Province / Territory by Level of Certification	Population Size	Total NGS Sample Size	In-scope Sample Size	Total NGS Respondents
Newfoundland and Labrador	4,275	2,652	2,595	2,009
Trade/vocational	248	248	247	200
College diploma	1,732	1,112	1,103	866
Bachelor's degree	1,959	956	920	722
Master's degree	309	309	299	212
Doctorate	27	27	26	9
Prince Edward Island	1,603	1,228	1,225	871
Trade/vocational	0	0	0	0
College diploma	1,070	811	811	597
Bachelor's degree	526	410	409	270
Master's degree	5	5	5	4
Doctorate	2	2	0	0
Nova Scotia	10,877	3,820	3,733	2,046
Trade/vocational	74	74	73	5
College diploma	4,185	1,619	1,575	761
Bachelor's degree	5,705	1,349	1,330	844
Master's degree	860	725	702	410
Doctorate	53	53	53	26
New Brunswick	6,706	3,054	2,766	1,790
Trade/vocational	415	415	165	110
College diploma	2,549	1,130	1,128	812
Bachelor's degree	3,286	1,070	1,047	659
Master's degree	420	403	390	195
Doctorate	36	36	36	14
Quebec	91,432	17,878	17,191	11,042
Trade/vocational	29,084	3,732	3,686	2,492
College diploma	14,465	2,084	2,066	1,520
Bachelor's degree	37,941	8,394	7,903	5,135
Master's degree	8,387	2,113	2,015	1,166
Doctorate	1,555	1,555	1,521	729
Ontario	123,036	9,882	9,631	6,324
Trade/vocational	4,791	2,334	2,242	1,393
College diploma	50,254	2,259	2,201	1,484
Bachelor's degree	57,058	1,917	1,865	1,293
Master's degree	9,478	1,917	1,885	1,272
Doctorate	1,455	1,455	1,438	882
Manitoba	8,275	4,522	4,144	3,009
Trade/vocational	767	732	485	333
College diploma	2,140	1,127	1,126	863
Bachelor's degree	4,796	2,091	1,981	1,432
Master's degree	484	484	468	330
Doctorate	88	88	84	51

Province / Territory by Level of Certification	Population Size	Total NGS Sample Size	In-scope Sample Size	Total NGS Respondents
Saskatchewan	8,995	4,297	4,076	2,800
Trade/vocational	1,205	1,057	937	566
College diploma	2,319	1,298	1,269	912
Bachelor's degree	4,767	1,276	1,214	874
Master's degree	609	571	564	398
Doctorate	95	95	92	50
Alberta	21,272	6,851	6,320	4,199
Trade/vocational	1,597	1,351	926	568
College diploma	5,583	2,221	2,179	1,516
Bachelor's degree	11,658	1,670	1,615	1,092
Master's degree	2,004	1,179	1,175	777
Doctorate	430	430	425	246
British Columbia	38,063	7,042	6,678	4,206
Trade/vocational	1,037	999	798	448
College diploma	18,024	2,314	2,191	1,370
Bachelor's degree	16,058	1,846	1,825	1,165
Master's degree	2,464	1,403	1,391	922
Doctorate	480	480	473	301
Yukon	143	116	110	75
Trade/vocational	106	106	100	69
College diploma	0	0	0	0
Bachelor's degree	37	10	10	6
Northwest Territories	198	198	173	105
Trade/vocational	115	115	91	51
College diploma	83	83	82	54
Nunavut	46	18	18	7
Trade/vocational	14	6	6	2
College diploma	30	10	10	5
Bachelor's degree	2	2	2	0
Canada	314,921	61,558	58,660	38,483
Trade/vocational	39,453	11,169	9,756	6,237
College diploma	102,434	16,068	15,741	10,760
Bachelor's degree	143,793	20,991	20,121	13,492
Master's degree	25,020	9,109	8,894	5,686
Doctorate	4,221	4,221	4,148	2,308

6.0 Data Collection

6.1 National Graduates Survey (Class of 2000)

A combination of self-study and classroom training was developed for this survey. Project supervisors from all the Statistics Canada Regional Offices came to head office for a two-day classroom training seminar. Interviewers conducted a self-study which involved reading the training manual, completing mock interviews on lap-top computers, answering review exercises and participating in a conference call to discuss any questions prior to the start of the survey.

Interviewers worked from their own homes and collected the data using a computer-assisted telephone interviewing method (CATI). They were instructed to make all reasonable attempts to obtain interviews with the selected graduates. Proxy response was not allowed. For graduates who at first refused to participate, a letter was sent from the Regional Office to the dwelling address stressing the importance of the survey and the graduates' cooperation. This was followed by a second call from the interviewer. For cases in which the timing of the interviewer's call was inconvenient, an appointment was arranged to call back at a more convenient time. For cases in which there was no one home, numerous call backs were made. If graduates had moved, various tracing methods were used to locate them.

The collection period was scheduled to run from the week of May 15th to July 31st, 2002. Collection was extended in some regions to allow interviewers to contact respondents and collect data up to August 23rd, 2002.

6.2 Follow-up of Graduates Survey (Class of 2000)

Project supervisors and Senior interviewers from all the Statistics Canada Regional Offices came to head office for a two-day classroom training seminar. Presentations on subject matter and methodology were made, along with mock interviews and a quiz/game. Project supervisors and Senior interviewers then conducted a 2-day training of interviewers in the Regional Offices, assisted with several on-line tutorials, mock interviews and the quiz/game.

Interviewers worked in the Regional Offices and collected the data using a computer-assisted telephone interviewing method (CATI). They were instructed to make all reasonable attempts to obtain interviews with the selected graduates. Proxy response was not allowed. For graduates who refused to participate, a letter was sent from the Regional Office to the dwelling address stressing the importance of the survey and the graduates' cooperation. This was followed by a second call from the interviewer. For cases in which the timing of the interviewer's call was inconvenient, an appointment was arranged to call back at a more convenient time. For cases in which there was no one home, numerous call backs were made. If graduates had moved, various tracing methods were used to locate them.

The collection period ran from April 27th, 2005 to July 24th, 2005.

7.0 Data Processing

This chapter presents a brief summary of the processing steps involved in producing the microdata file.

7.1 National Graduates Survey/Follow-up of Graduates (Class of 2000)

7.1.1 Data Capture

Responses to survey questions are captured directly by the interviewer at the time of the interview using a computerized questionnaire. The computerized questionnaire reduces processing time and costs associated with data entry, transcription errors, and data transmission. The response data are encrypted to ensure confidentiality and sent via modem to the appropriate Statistics Canada Regional Office. From there they are transmitted over a secure line to Ottawa for further processing.

Some editing is done directly at the time of the interview. Where the information entered is out of range (too large or small) of expected values, or inconsistent with previous entries, the interviewer is prompted, through message screens on the computer, to modify the information. However, for some questions interviewers have the option of bypassing the edits, and of skipping questions if the graduate does not know the answer or refuses to answer. Therefore, the response data are subjected to further edit and imputation processes once they arrive in head office.

7.1.2 Editing

The first stage of survey processing undertaken at head office was the replacement of any "out-of-range" values on the data file with blanks. This process was designed to make further editing easier.

The first type of error treated was errors in questionnaire flow, where questions which did not apply to the graduate (and should therefore not have been answered) were found to contain answers. In this case a computer edit automatically eliminated superfluous data by following the flow of the questionnaire implied by answers to previous, and in some cases, subsequent questions.

The second type of error treated involved a lack of information in questions which should have been answered. For this type of error, a non-response or "not-stated" code was assigned to the item.

7.1.3 Coding of Open-ended Questions

A few data items on the questionnaire were recorded by interviewers in an open-ended format. These were items relating to the type of education programs taken before and after graduation in 2000, as well as questions relating to the graduates industry and occupation. These open-ended questions were coded using various standard classifications (see Sections 7.1.3.1 and 7.1.3.2). An additional type of coding performed is called "Other – Specify" coding (see Section 7.1.3.3).

7.1.3.1 Coding of Education Programs

Field of study program descriptions were coded using the Classification of Instructional Programs (CIP – 2000, November 2001). Programs were coded at the six-digit level. See Appendix A for details on the code set.

7.1.3.2 Coding of Industry and Occupation

For each job held by the graduate in the reference periods, the questionnaire collected information on the name of the employer, the kind of business, industry or service the employer was in, the kind of work done and the usual duties or responsibilities of the graduate in the job. This information was used to assign industry and occupation codes to each job using the North American Industry Classification System (NAICS) 1997 and the National Occupational Classification for Statistics (NOC-S) 2001. See Appendices B and C for details on the code sets.

7.1.3.3 Coding of “Other – Specify” Answers

“Other – Specify” coding was done on questions that contained a list of answer categories that had “Other - Specify” as the final category. If the write-in was reflected in one of the existing categories, the response was recoded into the appropriate one. New categories may be added if there are a large number of write-ins which can be categorized together. Responses that cannot be coded into an existing category or into new categories are coded as “Other”.

7.1.4 Imputation

Imputation is the process that supplies valid values for those variables that have been identified for a change either because of invalid information or because of missing information. The new values are supplied in such a way as to preserve the underlying structure of the data and to ensure that the resulting records will pass all required edits. In other words, the objective is not to reproduce the true microdata values, but rather to establish internally consistent data records that yield good aggregate estimates.

We can distinguish between three types of non-response. Complete non-response is when the graduate does not provide the minimum set of questions. These records are dropped and accounted for in the weighting process (see Chapter 9.0). Item non-response is when the graduate does not provide an answer to one question, but goes on to the next question. These are usually handled using the “not stated” code or are imputed. Finally, partial non-response is when the graduate provides the minimum set of questions but does not finish the interview. These records can be handled like either complete non-response or multiple item non-response.

For quantitative variables such as financial variables, editing which includes outlier detection and imputation was performed. These variables include reported information on personal income and student loans. Reported values were grouped based on field of study, level of certification and preferred mode of reporting the data (i.e., hourly, daily, weekly, yearly, etc.). Potential outliers were identified using several statistical methods. Manual investigations were then made on these cases to confirm their outlier status. Outliers were replaced by a more plausible value, or coded to not stated. The latter is the only imputation that was performed for the Follow-up of Graduates Survey – Class of 2000. Further information on the variables which were imputed for the National Graduates Survey – Class of 2000 can be found in Chapter 9 of the following document: *Micro Data User Guide – National Graduates Survey – Class of 2000*.

7.1.5 *Creation of Derived Variables*

Combining Items

A number of variables have been derived by combining questions on the questionnaire in order to facilitate data analysis. For example, six questions from the Activities Last Week (LF) section are used to derive labour force status in the week prior to the interview (LFSTAT). These included:

LF_Q02 - [Last week], were you enrolled full-time or part-time [in any credit courses at an educational or training institution]?

LF_Q03 - Last week, did you work at a job or a business?

LF_Q05 - Were you absent from work [last week] because of a temporary layoff?

LF_Q07 - Last week, did you have a job to start at a definite date in the future?

LF_Q10 - Last week, were you looking for a job?

LF_Q11 - [Last week], were you looking for a job at which you would usually work 30 or more hours per week?

For a list of the derived variables available on the PUMF and a description of how they were derived, see Appendix D.

8.0 Response Rates

This chapter describes the response rates for the Follow-Up of Graduates – Class of 2000 (FOG2000). Survey response rates are measures of the effectiveness of the population being sampled and the collection process. They are also a good indicator of the quality of the estimates produced.

A respondent is a person for whom there is usable minimal information on the questionnaire. Cases where the graduates did not go far enough in the questionnaire or where crucial questions (e.g. diploma or degree obtained, employment status) were not answered, were deemed non-responding units.

The overall response rate for the FOG2000 is 68.5%. However, it should be reminded that FOG's sample is comprised of NGS' respondents. Details on NGS2000 response rates can be found in the following document: *Micro Data User Guide – National Graduates Survey – Class of 2000*.

The following table presents the collection results for the FOG2000. The final sample size was 34,304, which represents all of NGS2000 respondents minus trade/vocational graduates based on the reported certification level variable.

Please note that, due to some inconsistencies between the stratification variable and the reported certification level, a small number of respondents other than trade/vocational graduates were not included in the FOG2000 sample, thus, leading to a smaller response rate in some provinces and certification levels.

Response Rates by Province / Territory and Level of Certification – Unweighted

Province / Territory by Level of Certification	FOG Sample Size	Responding Graduates	Response Rate (%)
Newfoundland and Labrador	1,845	1,221	66.2
College	904	530	58.6
Bachelor's Degree	703	494	70.3
Master	228	189	82.9
Doctorate	10	8	80.0
Prince Edward Island	830	617	74.3
College	552	399	72.3
Bachelor's Degree	263	205	77.9
Master	9	9	100.0
Doctorate	6	4	66.7
Nova Scotia	1,982	1,512	76.3
College	696	512	73.6
Bachelor's Degree	828	620	74.9
Master	430	359	83.5
Doctorate	28	21	75.0
New Brunswick	1,761	1,205	68.4
College	897	549	61.2
Bachelor's Degree	652	485	74.4
Master	198	162	81.8
Doctorate	14	9	64.3

Province / Territory by Level of Certification	FOG Sample Size	Responding Graduates	Response Rate (%)
Quebec	8,565	6,445	75.2
College	1,634	1,206	73.8
Bachelor's Degree	5,017	3,776	75.3
Master	1,260	976	77.5
Doctorate	654	487	74.5
Ontario	6,105	3 304	54.1
College	2,658	908	34.2
Bachelor's Degree	1,264	776	61.4
Master	1,304	917	70.3
Doctorate	879	703	80.0
Manitoba	2,911	1 985	68.2
College	1,100	621	56.5
Bachelor's Degree	1,397	1 021	73.1
Master	361	292	80.9
Doctorate	53	51	96.2
Saskatchewan	2,450	1 693	69.1
College	1,112	612	55.0
Bachelor's Degree	877	711	81.1
Master	410	332	81.0
Doctorate	51	38	74.5
Alberta	3,783	2,699	71.3
College	1,669	1,016	60.9
Bachelor's Degree	1,123	883	78.6
Master	743	608	81.8
Doctorate	248	192	77.4
British Columbia	3,903	2,756	70.6
College	1,406	879	62.5
Bachelor's Degree	1,225	897	73.2
Master	973	750	77.1
Doctorate	299	230	76.9
Yukon	73	5	6.8
College	70	3	4.3
Bachelor's Degree	3	2	66.7
Northwest Territories	90	42	46.7
College	89	41	46.1
Bachelor's Degree	1	1	100.0
Nunavut	6	4	66.7
College	6	4	66.7
Canada	34,304	23,488	68.5
College	12,793	7,280	56.9
Bachelor's Degree	13,353	9,871	73.9
Master	5,916	4,594	77.7
Doctorate	2,242	1,743	77.7

A subsample of the FOG master file, consisting of 11,200 records, was selected for the PUMF. For confidentiality reasons, a provincial breakdown of records on the PUMF cannot be provided.

9.0 Weighting and Sampling Variability

In order for estimates produced from survey data to be representative of the target population, and not just of the sample itself, users must incorporate the survey weights into their calculations. A survey weight is given to each person included in the final sample, that is, the sample of persons who responded to the survey questions. This weight corresponds to the number of persons represented by the respondent for the target population. If the frame used was perfect (covering exactly the population of interest) and all selected units were traced, contacted and completed the survey, then the design weight assigned to each unit, given by the inverse of the probability of selection of each unit in the sample, would represent accurately and exactly the number of graduates in the target population. In this situation, using this weight would yield unbiased estimates. However, this is not the case when surveys are faced with non-response and imperfect frames. Weight adjustments are traditionally used to compensate for these different issues. Response patterns have to be studied carefully to appropriately correct for non-response by creating response homogeneous groups (RHG) based on the characteristics of the respondents and the non-respondents.

For weighting purposes and in order to facilitate the variance estimation, the FOG2000 can be seen as a three-phase survey. The first phase corresponds to the selection of the NGS2000 sample and the NGS responding units correspond to the second phase sample. The underlying assumption is that the second phase sample is a sub-sample of the first phase sample. Note that in practice, the second phase is a Bernoulli sample and the second phase sampling probabilities are equal to the observed response probabilities in the RHGs. More details can be found about this two-phase model for non-response in Särndal, Swensson and Wretman.¹ Similarly, the FOG2000 responding units correspond to the third phase sample and the third phase sampling probabilities are equal to the observed response probabilities in another set of RHGs specifically determined for the FOG2000 non-response.

As indicated in Section 1, the FOG2000 PUMF represents a sub-sample of the FOG2000 Master File. However, instead of adding a 4th phase to the weighting process (by subsampling the FOG2000 respondents), the subsampling step was conducted within the 1st phase of the survey (NGS sample selection). That is, the original NGS sample was reduced within each design stratum in such a way as to reduce the risk of disclosure. Therefore, the 1st phase weights were recalculated based on the new sample size in each stratum. The 2nd and 3rd phase adjustments were computed in the same way as for the FOG2000 Master File.

The following section describes in details the weighting strategy used for the FOG2000 PUMF.

NGS Design Weight – PUMF version (phase 1)

At the time of selection, an initial design weight was assigned to each graduate, as the inverse of its probability of selection. Since the NGS2000 design is stratified with simple random sampling within strata, the probability of selection of the graduate i in stratum h is:

$$\pi_{ih}^{phase1} = \frac{n_h}{N_h}$$

where, n_h and N_h denote respectively the sample (i.e., the sub-sample for the PUMF) and population size of stratum h . The design weight is given by $\frac{1}{\pi_{ih}^{phase1}}$

¹ Särndal, C.E., B. Swensson and J. Wretman 1992. Model Assisted Survey Sampling. Springer-Verlag, New York.

NGS Non-response Adjustment (phase 2)

A non-response adjustment was also applied based on RHGs. RHGs were developed with the premise of identifying sample units with similar response probabilities. In other words, it is assumed that graduates pertaining to a given RHG are equally likely to respond to the survey in a similar fashion. Analyses were completed and the RHGs were identified. For the NGS non-response adjustment, those RHGs were formed using stratification variables, although they do not perfectly correspond to strata given aggregations made to ensure a sufficient size by group.

For graduate i in RHG j the response probability is calculated as:

$$\pi_{ij}^{phase2} = \frac{\text{number of responding units in RGH } j}{\text{number of sample units in RGH } j}$$

The phase-2 weight is given by $\frac{1}{\pi_{ij}^{phase2}}$

FOG Non-response Adjustment (phase 3)

The FOG2000 can be considered as a third phase of the NGS2000 and similarly to the previous adjustment, a non-response adjustment was also applied based on RHGs to account for the FOG2000 non-response. However, more information on the non-respondents is available for this second round of non-response since both FOG2000 respondents and non-respondents responded to the NGS2000. For this reason, NGS2000 variables were used to create more precise RHGs. The RHGs were created using the approach proposed by Eltinge and Yansaneh.¹ This approach consists in finding RHGs using response probabilities calculated using a logistic regression model and by grouping together the graduates with the same probability of response.

Logistic regression is a statistical process by which one attempts to estimate the value of a binary outcome value given a combination of auxiliary information. In the case of FOG, the value of some sort of response event (1=response, 0=non-response) is modeled based on a set of variables from the NGS frame and data set.

Once again, for graduate i in RHG k the response probability is calculated as:

$$\pi_{ik}^{phase3} = \frac{\text{number of responding units in RGH } k}{\text{number of sample units in RGH } k}$$

The phase-3 weight is given by $\frac{1}{\pi_{ik}^{phase3}}$

Post-Stratification

Post-stratification is one of the calibration estimation techniques widely used in social surveys. It allows benchmarking on population counts. Note that the post-stratification file still represents the target population and the FOG2000 used the same post-stratification file created for the NGS2000 since the target population is the same for both cycles. As for the NGS2000, post-stratification classes were created based on the province, the level of certification and the field of study. After merging some classes showing low counts, 143 post-stratification classes were created.

The post-stratification adjustment is calculated at the post-stratum level using the following formula:

¹ Eltinge, J. and Yansaneh, I. 1997. Diagnostics for Formation of Nonresponse Adjustment Cells, With an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey, Survey Methodology, June 1997.

$$w_{i, ps} = \frac{\text{population estimate for a given post - stratum}}{\text{sum of the weights of respondents in a given post - stratum}}$$

FOG Final Weight

Consequently, the final weight for graduate i is formed by multiplying the weights obtained from the three phases of the FOG2000, along with the post-stratification adjustment. The final weight is given by:

$$w_i = w_{ih} \times w_{ij} \times w_{ik} \times w_{i, ps}$$

10.0 Data Quality

This chapter provides the user with information about the various factors affecting the quality of the survey data. There are two main types of errors: sampling errors and non-sampling errors. A sampling error is the difference between an estimate derived from a sample and the one that would have been obtained from a census that used the same procedures to collect data from every person in the population. All other types of errors such as frame coverage, response, processing and non-response are non-sampling errors. Many of these errors are difficult to identify and quantify. These are discussed in Section 10.2.

10.1 Sampling Errors

The estimates derived from the Follow-Up of Graduates – Class of 2000 (FOG2000) are based on a sample of graduates and not from a complete enumeration (census). This difference is the sampling error of the estimates.

The basis for measuring sampling error 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 measure, known as the coefficient of variation (CV) of an estimate, is obtained by expressing the standard error of the estimate as a percentage of the estimate. This measure allows for better comparisons of quality between different types of estimates. The smaller the CV, the smaller the sampling variability, meaning smaller CVs are more desirable. The CV depends on the size of the sample on which the estimate is based, the population size and on the distribution of the sample, i.e. the sampling fraction of the units of the domains being estimated. The table in Section 11.5 presents the characteristics of some CVs and the Statistics Canada guidelines for release.

Note that for the FOG2000, the error due to non-response has been incorporated into the sampling error. The use of the Generalized Estimation System (GES) takes into account the non-response variability into the estimates variability. Refer to section 11.4 for more information on variance estimation and guidelines for statistical analysis.

10.2 Non-sampling Errors

There are many sources of non-sampling errors that are not related to sampling, but may occur at almost any phase of a survey operation. Interviewers may misunderstand survey instructions, graduates may make a mistake in answering the questions, responses may be recorded in the questionnaire incorrectly or errors may be made in the processing or tabulating of the data. For the FOG2000, quality assurance measures were implemented at each phase of the data collection to monitor the quality of the data. These measures included precise interviewer training with respect to the survey procedures and questionnaire, observation of interviews to detect questionnaire design problems or misinterpretation of instructions and coding and edit quality checks to verify the processing logic. Chapter 7.0 outlines data processing procedures. Other kinds of non-sampling error are more easily quantifiable, especially non-response and population frame under-/over-coverage, the topics of the next two sections.

10.3 Non-response

Non-response, if not appropriately corrected, is a type of error that can lead to bias in the survey estimates. For the FOG2000, non-response reduced significantly the number of usable records, along with non-response from the NGS2000, given that the FOG sample is a sub-sample of the NGS respondents. Biased estimates can occur when unusable units have significantly different characteristics from the usable ones. In Chapter 8.0, non-response rates were computed for basic domains to describe its extent. Extensive studies were completed on non-response to construct the proper adjustment weights for the FOG2000. Since the use of the final weights will yield the appropriate estimates of the population counts and ensure that non-respondents are incorporated and accounted for, it stresses the importance of using the final weights in any tabulations or analysis using the FOG2000 data. Any estimation done without the use of weights may produce biased or incorrect results.

10.4 Coverage

Coverage is an indication of how a survey frame covers the target population. There could be over-coverage if the survey frame contains units that should not have been included, such as deaths, duplicates, or incorrect date of graduation captured on the file. There could also be under-coverage, if the survey frame missed some units that should have been included.

For the NGS2000, there was some under-coverage for graduates of colleges in southern Alberta. Data required to build the frame could not be obtained from these institutions. They were not covered on the frame. Consequently, they could not be selected nor represented in any tabulation. This problem also affects the FOG2000.

**Summary of Institutions Reporting Graduates for the National Graduates Survey –
Class of 2000 Population Frame**

Province / Territory	Institutions			Missing	
	Received	Expected	%	Colleges	Universities
Newfoundland and Labrador	2	2	100	0	0
Prince Edward Island	3	3	100	0	0
Nova Scotia	15	15	100	0	0
New Brunswick	10	10	100	0	0
Quebec	319	319	100	0	0
Ontario	63	63	100	0	0
Manitoba	11	12	92	1	0
Saskatchewan	11	11	100	0	0
Alberta	19	28	68	9	0
British Columbia	26	27	96	1	0
Yukon	1	1	100	0	0
Northwest Territories	1	1	100	0	0
Nunavut	1	1	100	0	0
Canada	482	493	98	11	0

Note: Affiliated institutions are not always reported as separate institutions. The number of institutions also excludes those with no graduates in the calendar year 2000.

11.0 Guidelines for Tabulation Analysis and Release

This chapter of the documentation outlines 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 Statistics Canada and, at the same time, will be able to develop currently unpublished figures in a manner consistent with these established guidelines.

11.1 Rounding Guidelines

In order that estimates for publication or other release derived from the Follow-Up of Graduates – Class of 2000 (FOG2000) microdata file 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 ratio) 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.

11.2 Sample Weighting Guidelines for Tabulation

The FOG2000 uses a stratified simple random sample design without replacement of graduates within strata. When producing simple estimates, including the production of ordinary statistical tables, users must use the final weight associated with the graduates concerned by the analysis. If final weights are not used, the estimates derived from the microdata file cannot be considered to be representative of the survey population and will not correspond to those produced by Statistics Canada. The final weight assigned to a given responding graduate reflects the number of graduates in the FOG2000's population he/she represents.

For any analysis dealing with correlation analysis or any other statistics where a significance measure is required, it is recommended that an adjusted weight be used. This weight is obtained by multiplying the final weight by the sample size and dividing this total by the total estimated population. This produces a mean weight of 1 and a sum of weights equal to the sample size.

The benefit of this adjusted weight is that an overestimation of the significance (which is very sensitive to sample size) is avoided while maintaining the same distributions as those obtained when using the demographic weight. The disadvantage is that the numerator is not weighted up to the target population and the coefficient of variance is no longer useful as a measure of data quality.

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.

11.3 Definitions of Types of Estimates: Categorical and Quantitative

The PUMF has been set up so that the graduate is the unit of analysis. The final weight that can be found on each record is called FWTPP in the codebook.

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 or the proportion of self-employed graduates working at a job last week is an example 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: Last week, did you work at a job or a business?
R: Yes / No

Q: At your (main) job last week, were you a paid worker or self-employed?
R: Paid worker / Self-employed / Unpaid family worker

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.

An example of a quantitative estimate is the average number of hours worked per week at a job. The numerator is an estimate of the total number of hours worked per week and its denominator is the number of graduates working.

Examples of Quantitative Questions:

Q: How many (paid) hours a week do you usually work at this job?

R: |_|_|_| hours

Q: How much do you now owe for all your government-sponsored student loans?

R: |_|_|_|_|_| dollars

11.3.1 Tabulation of Categorical Estimates

Estimates of the number of graduates 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:

- summing the final weights of records having the characteristic of interest for the numerator (\hat{X}),
- summing the final weights of records having the characteristic of interest for the denominator (\hat{Y}), then
- dividing estimate a) by estimate b) (\hat{X}/\hat{Y}).

11.3.2 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 by graduates in their main job in the week before they were surveyed multiply the value reported in question FLFQ79 (hours worked per week) by the final weight for the record, then sum this value over all records with LFSTAT05 = 1 (employed) and FLFQ79 < 996.

To obtain a weighted average of the form \hat{X}/\hat{Y} , the numerator (\hat{X}) is calculated as for a quantitative estimate and the denominator (\hat{Y}) is calculated as for a categorical estimate. For example, to estimate the average number of hours worked by graduates in their main job in the week before they were surveyed,

- estimate the total number of hours (\hat{X}) as described above,
- estimate the number of graduates (\hat{Y}) in this category by summing the final weights of all records with LFSTAT05 = 1 and FLFQ79 < 996, then
- divide estimate a) by estimate b) (\hat{X}/\hat{Y}).

11.4 Guidelines for Statistical Analysis

The FOG2000 is based upon a complex design, with stratification, multiple phases 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 can differ from what 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 the FOG2000 PUMF, approximate release cut-offs have been calculated and are presented in Section 11.6. As well, approximate variances for simple estimates such as totals, proportions and ratios (for qualitative variables) can be derived using the accompanying Approximate Sampling Variability Tables (see Appendix E).

For many analysis techniques (for example linear regression, logistic regression, analysis of variance), a method exists that 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 the multiple phases of the sample's design, but they will take into account the unequal probabilities of selection. The rescaling can be accomplished by using in the analysis a weight equal to the original weight divided by the average of the original weights for the sampled units (people) contributing to the estimator in question.

11.5 Release Guidelines

Before releasing and/or publishing any estimate from the FOG2000, users should first determine 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 10.0.

First, the number of graduates (unweighted) who contribute to the calculation of the estimate should be determined. If this number is less than 30, the weighted estimate should be considered of unacceptable quality and more importantly too small for disclosure. Users are invited to read the document Statistics Canada Quality Guidelines available on Statistics Canada web site.

Once this condition is met, users must determine the coefficient of variation of the estimate and follow the guidelines below. All estimates can be considered releasable. However, those of marginal or unacceptable quality level must be accompanied by a warning to caution subsequent users. These quality level guidelines should be applied to weighted rounded estimates.

Quality Level Guidelines

Quality Level of Estimate	Guidelines
1) Acceptable	<p>Estimates must have <u>all</u> of the following characteristics: a sample size of 30 graduates or more, and low coefficients of variation in the range of 0.0% to 16.5%.</p> <p>No warning is required.</p>
2) Marginal	<p>Estimates must have <u>all</u> of the following characteristics: a sample size of 30 graduates or more, and high coefficients of variation in the range of 16.6% to 33.3%.</p> <p>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.</p>
3) Unacceptable	<p>Estimates must have <u>at least one</u> of the following characteristics: a sample size of less than 30 graduates, or very high coefficients of variation in excess of 33.3%.</p> <p>Statistics Canada recommends not to release estimates of unacceptable quality. However, if the user chooses to do so then estimates should be flagged with the letter U (or some similar identifier) and the following warning should accompany the estimates:</p> <p>“Please be warned that these estimates [flagged with the letter U] do not meet Statistics Canada’s quality standards. Conclusions based on these data will be unreliable, and most likely invalid.”</p>

11.6 Release Cut-off's for the PUMF

The following table provides an indication of the precision of population estimates as it shows the release cut-offs associated with a CV of 16.5% and a CV of 33.3% (correspond to quality levels presented in the previous section). These cut-offs are derived from the coefficient of variation (CV) tables discussed in Chapter 12.0. For example, the table shows that the quality of a weighted estimate of 500 college level graduates possessing a given characteristic is marginal-

Note that these cut-offs apply to estimates of population totals only. To estimate ratios, users should not use the numerator value (nor the denominator) in order to find the corresponding quality level. Rule 4 in Section 12.1 and Example 4 in Section 12.1.1 explain the correct procedure to be used for ratios.

Domain	CV of 16.5% Min X	CV of 33.3% Min X
Canada (all respondents)	1,515	375
College Level (CERTLEVP=1)	1,465	365
Bachelor Level (CERTLEVP=2)	1,700	425
Master/Doctorate Level (CERTLEVP=3)	900	230

12.0 Approximate Sampling Variability Tables

In order to supply coefficients of variation (CV) that 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 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 are derived using the variance formula for simple random sampling, and incorporating a factor which reflects the sample design and the adjustment for nonresponse. 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 (usually the 75th percentile) to be used in the CV tables, which would then apply to the entire set of characteristics.

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. Since the approximate CV is conservative, the use of actual variance estimates may cause the estimate to be switched from one quality level to another. For instance a *marginal* estimate could become *acceptable* based on the exact CV calculation.

Remember: If the number of observations on which an estimate is based is less than 30, the weighted estimate is most likely unacceptable and Statistics Canada recommends not releasing such an estimate, regardless of the value of the coefficient of variation.

12.1 How to Use the Coefficient of Variation Tables for Categorical Estimates

The following rules should enable the user to determine the approximate coefficients of variation from the Approximate 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 of Persons Possessing a Characteristic (Aggregates)

The coefficient of variation depends only on the size of the estimate itself. On the Approximate Sampling Variability Table for the appropriate level of certification, 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 of Persons 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 working persons who are self-employed is more reliable than the estimated number of self-employed persons. (Note that in the tables the coefficients of variation decline in value reading from left to right).

When the proportion or percentage is based upon the total population 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 $\sigma_{\hat{d}}/\hat{d}$. 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 working persons and the numerator is the number of self-employed persons.

In cases where the numerator is not a subset of the denominator, for example, the ratio of the number of self-employed males as compared to the number of self-employed females, the standard error 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 \hat{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 $\sigma_{\hat{R}}/\hat{R}$. 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 CVs for the two ratios are first determined using Rule 4, and then the CV of their difference is found using Rule 3.

12.1.1 Examples of Using the Coefficient of Variation Tables for Categorical Estimates

The following examples based on the FOG2000 PUMF are included to assist users in applying the above rules.

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

Suppose that a user estimates that 24,591 graduates had difficulties repaying their student loans. How does the user determine the coefficient of variation of this estimate?

- 1) Refer to the coefficient of variation table for Canada.
- 2) The estimated aggregate (24,591) 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 25,000.
- 3) The coefficient of variation for an estimated aggregate is found by referring to the first non-asterisk entry on that row, in this case 3.9%.
- 4) So the approximate coefficient of variation of the estimate is 3.9%. The finding that there were 24,591 graduates (to be rounded according to the rounding guidelines in Section 11.1) who had difficulties repaying their student loans is publishable with no qualifications.

Example 2: Estimates of Proportions or Percentages of Persons Possessing a Characteristic

Suppose that the user estimates that $11,745 / 24,591 = 47.8\%$ of graduates who had difficulties repaying their student loans are married or in common-law relationships. How does the user determine the coefficient of variation of this estimate?

- 1) Refer to the coefficient of variation table for Canada.

Because the estimate is a percentage based on a subset of the total population (i.e., graduates who had difficulties repaying their student loans), it is necessary to use both the percentage (47.8%) and the numerator portion of the percentage (11,745) in determining the coefficient of variation.

- 2) The numerator, 11,745, 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 10,000. Similarly, the percentage estimate does not appear as any of the column headings, so it is necessary to use the percentage closest to it, 50.0%.
- 3) The figure at the intersection of the row and column, 4.5%, is the coefficient of variation to be used.
- 4) So the approximate coefficient of variation of the estimate is 4.5%. The finding that 47.8% of graduates who had difficulties repaying their student loans are married or in common-law relationships can be published with no qualifications.

Example 3: Estimates of Differences Between Aggregates or Percentages

Suppose that a user estimates that $4,312 / 9,548 = 45.2\%$ of male graduates who had difficulties repaying their student loans are married or in common-law relationships, while $7,433 / 15,043 = 49.4\%$ of female graduates who had difficulties repaying their student loans are married or common-law. How does the user determine the coefficient of variation of the difference between these two estimates?

- 1) Using the Canada coefficient of variation table in the same manner as described in Example 2 gives the CV of the estimate for men as 7.2%, and the CV of the estimate for women as 5.4%.

- 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 (women), \hat{X}_2 is estimate 2 (men), 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 $\hat{d} = 0.494 - 0.452 = 0.042$ is:

$$\begin{aligned}\sigma_{\hat{d}} &= \sqrt{[(0.494)(0.054)]^2 + [(0.452)(0.072)]^2} \\ &= \sqrt{(0.000712) + (0.001059)} \\ &= 0.042\end{aligned}$$

- 3) The coefficient of variation of \hat{d} is given by $\sigma_{\hat{d}} / \hat{d} = 0.042 / 0.042 = 1.000$
- 4) So the approximate coefficient of variation of the difference between the estimates is 100.0%. The difference between the estimates is considered unacceptable and Statistics Canada recommends this estimate not be released. However, should the user choose to do so, the estimate should be flagged with the letter U (or some similar identifier) and be accompanied by a warning to caution subsequent users about the high levels of error associated with the estimate.

Example 4: Estimates of Ratios

Suppose that the user estimates that 35,826 males supervised other employees at their main job last week, while 40,359 females supervised other employees at their main job last week. The user is interested in comparing the estimate of men versus women 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 male graduates who supervised other employees at their main job last week. The denominator of the estimate (\hat{X}_2) is the number of female graduates who supervised other employees at their main job last week.
- 2) Refer to the coefficient of variation table for Canada.
- 3) The numerator of this ratio estimate is 35,826. The figure closest to it is 40,000. The coefficient of variation for this estimate is found by referring to the first non-asterisk entry on that row, namely 3.0%.
- 4) The denominator of this ratio estimate is 40,359. The figure closest to it is 40,000. The coefficient of variation for this estimate is found by referring to the first non-asterisk entry on that row, 3.0%
- 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:

$$\begin{aligned}\alpha_{\hat{R}} &= \sqrt{(0.03)^2 + (0.03)^2} \\ &= \sqrt{0.0009 + 0.0009} \\ &= 0.042\end{aligned}$$

- 6) The obtained ratio of male graduates versus female graduates who supervised other employees at their main job last week is 35,826 / 40,359, which is 0.89 (to be rounded according to the rounding guidelines in Section 11.1). The coefficient of variation of this estimate is 4.2%, which makes the estimate releasable with no qualifications.

Example 5: Estimates of Differences of Ratios

Suppose that the user estimates that the ratio of male to female graduates who supervised other employees at their main job last week, is 0.96 at the College certification level (CERTLEVP=1) and 0.85 at the University level (CERTLEVP=2 or 3). The user is interested in comparing the two ratios to see if there is a statistical difference between them. How does the user determine the coefficient of variation of the difference?

- 1) First calculate the approximate coefficient of variation for the College ratio (\hat{R}_1) and the University ratio (\hat{R}_2) as in Example 4. The approximate CV for the College ratio is 7.1%, and 5.5% for the University ratio.
- 2) Using Rule 3, the standard error of a difference ($\hat{d} = \hat{R}_1 - \hat{R}_2$) is:

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

where α_1 and α_2 are the coefficients of variation of \hat{R}_1 and \hat{R}_2 respectively. That is, the standard error of the difference $\hat{d} = 0.85 - 0.96 = -0.11$ is:

$$\begin{aligned}\sigma_{\hat{d}} &= \sqrt{[(0.85)(0.055)]^2 + [(0.96)(0.071)]^2} \\ &= \sqrt{(0.002186) + (0.004646)} \\ &= 0.083\end{aligned}$$

- 3) The coefficient of variation of \hat{d} is given by $\sigma_{\hat{d}} / \hat{d} = 0.083 / -0.11 = -0.755$.
- 4) So the approximate coefficient of variation of the difference between the estimates is 75.5%. The difference between the estimates is considered unacceptable and Statistics Canada recommends this estimate not be released. However, should the user choose to do so, the estimate should be flagged with the letter U (or some similar identifier) and be accompanied by a warning to caution subsequent users about the high levels of error associated with the estimate.

12.2 How to Use the Coefficient of Variation 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, 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 under the assumption 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 of 100 that the difference 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 ($CI_{\hat{X}}$):

$$CI_{\hat{X}} = (\hat{X} - t\hat{X}\alpha_{\hat{X}}, \hat{X} + t\hat{X}\alpha_{\hat{X}})$$

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;
- $t = 2.6$ 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.

12.2.1 Example of Using the Coefficient of Variation Tables to Obtain Confidence Limits

A 95% confidence interval for the estimated proportion of graduates who are married or in common-law relationships among those who had difficulties repaying their student loans (from Example 2, Section 12.1.1) would be calculated as follows:

$$\hat{X} = 47.8\% \text{ (or expressed as a proportion 0.478)}$$

$$t = 2$$

$$\alpha_{\hat{x}} = 4.5\% \text{ (0.045 expressed as a proportion) is the coefficient of variation of this estimate as determined from the tables.}$$

$$CI_{\hat{x}} = \{0.478 - (2) (0.478) (0.045), 0.478 + (2) (0.478) (0.045)\}$$

$$CI_{\hat{x}} = \{0.478 - 0.043, 0.478 + 0.043\}$$

$$CI_{\hat{x}} = \{0.435, 0.521\}$$

With 95% confidence, it can be said that between 43.5% and 52.1% of graduates who have difficulties repaying their student loans are married or in common-law relationships.

12.3 How to Use the Coefficient of Variation 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 \hat{X}_1 and \hat{X}_2 be sample estimates for two characteristics of interest. Let the standard error on the difference $\hat{X}_1 - \hat{X}_2$ be $\sigma_{\hat{d}}$.

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. In other words, the difference between the estimates is significant.

12.3.1 Example of Using the Coefficient of Variation Tables to Do a T-test.

Let us suppose that the user wishes to test, at 5% level of significance, the hypothesis that there is no difference between the proportion of male and female graduates who are married or in common-law relationships among those who had difficulties repaying their student loans. From Example 3, Section 12.1.1, the standard error of the difference between these two estimates was found to be 0.042. Hence,

$$t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_d} = \frac{0.494 - 0.452}{0.042} = \frac{0.042}{0.042} = 1.00$$

Since $t = 1.00$ is between -2 and 2, then no conclusion about the difference between the characteristics is justified at the 5% level of significance.

12.4 Coefficients of Variation for Quantitative Estimates

Special tables would have to be produced to determine the sampling error of quantitative estimates. Since most of the variables for the PUMF are primarily categorical in nature, this has not been done.

12.5 Coefficient of Variation Tables

Approximate Sampling Variability Tables are available in Appendix E.

13.0 Questionnaire, Code Sheets and Description of Derived Variables

Please refer to the files listed below for the Follow-up of Graduates Survey – Class of 2000 (FOG2000).

Questionnaire:

NGS2000_QuestE.doc
NGS2000_QuestE.pdf

FOG2000_QuestE.doc
FOG2000_QuestE.pdf

Appendices:

Classification of Instructional Programs (CIP)

Appendix A – CIP Aggregate_pumf.doc
Appendix A - CIP Aggregate_pumf.pdf

North American Industry Classification System (NAICS) 1997

Appendix B – NAICS_pumf.doc
Appendix B – NAICS_pumf.pdf

National Occupational Classification for Statistics (NOC-S) 2001

Appendix C – NOC-S_pumf.doc
Appendix C – NOC-S_pumf.pdf

Description of Derived Variables available on the PUMF

Appendix D – Documentation of Derived Variables_pumf.doc
Appendix D – Documentation of Derived Variables_pumf.pdf

Approximate Sampling Variability Tables

Appendix E – CV_Tables_pumf.doc
Appendix E – CV_Tables_pumf.pdf

Field of Study

Appendix F - Field of Study_pumf.doc
Appendix F - Field of Study_pumf.pdf

14.0 Record Layout with Univariate Frequencies

See FOG2000_PUMF_CODEBOOK.pdf or FOG2000_PUMF_CODEBOOK.doc for the record layout with univariate counts for the public use microdata file.