Social and Aboriginal Statistics Division

General Social Survey

Cycle 30: Canadians at Work and Home

Public Use Microdata File Documentation and User's Guide

Catalogue no. 12M0030X

June 2018
Aussi disponible en français

Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Social and Aboriginal Statistics Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (E-mail: statcan.sasdclientservices-dsseaserviceaclientele.statcan@canada.ca).

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.gc.ca or contact us by e-mail at STATCAN.infostats-infostats.STATCAN@canada.ca or by telephone from 8:30 a.m. to 4:30 p.m. Monday to Friday:

Statistics Canada National Contact Centre

Toll-free telephone (Canada and the United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-514-283-8300
Fax line	1-613-951-0581

Depository services program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

Accessing and ordering information

This product, Catalogue no. 12M0030X is also available as a standard printed publication.

The printed version of this publication can be ordered by:

• Telephone (Canada and United States)	1-800-267-6677
• Fax (Canada and United States)	1-877-287-4369
• E-mail	STATCAN.infostats-infostats.STATCAN@canada.ca
• Mail	Statistics Canada R.H. Coats Bldg., 6th Floor 150 Tunney's Pasture Driveway Ottawa, Ontario K1A 0T6

• In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

_____

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1-800-263-1136. The service standards are also published on www.statcan.ca under “About us” > “Providing services to Canadians.”

Statistics Canada
Social and Aboriginal Statistics Division

General Social Survey
Cycle 30: Canadians at Work and Home

Public Use Microdata File Documentation and User's Guide

By Anna Kemeny

Published by authority of the Minister responsible for Statistics Canada

June 2018

Catalogue no. 12M0030X

Frequency: Occasional

Ottawa

Cette publication est aussi disponible en français (no 12M0030X au catalogue)

Note of appreciation
Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

2016 GSS: Canadians at Work and Home

User Guide for the Public Use Microdata File (PUMF)

Table of Contents

1. Introduction

This guide was prepared for users of the Public Use Microdata File (PUMF) of the 2016 General Social Survey (GSS) on *Canadians at Work and Home*. Its objectives are to provide context and background information, to familiarize users with the content of the survey, and to describe procedures and concepts related to data quality, estimation, collection, processing and methodology.

The 2016 GSS, conducted from August 2nd to December 23rd 2016, is a sample survey with cross-sectional design. The target population includes all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada. The survey uses a new frame, created in 2013, that combines telephone numbers (landline and cellular) with Statistics Canada's Address Register, and collects data both electronically and via telephone. Data are subject to both sampling and non-sampling errors.

The information in the following sections should be used to ensure a clear understanding of the basic concepts that define the data provided in the GSS Cycle 30 PUMF, the underlying methodology of the survey and the key aspects of data quality. This information will provide a better understanding of the strengths and limitations of the data, and how they can be effectively used and analyzed. The information may be of particular importance when making comparisons with data from other surveys or sources of information, and in drawing conclusions regarding change over time, or differences between sub-groups of the target population.

2. Objectives of the General Social Survey

The GSS program, established in 1985, conducts telephone surveys across the ten provinces. The GSS is recognized for its regular collection of cross-sectional data that allows for trend analysis, and its capacity to test and develop new concepts that address current or emerging issues.

The two primary objectives of the General Social Survey are:

a) To gather data on social trends in order to monitor changes in the living conditions and well-being of Canadians over time; and
b) To provide information on specific social policy issues of current or emerging interest.

To meet these objectives, data collected by the GSS comprise two components: core content and classification variables. Core content is designed to measure changes in society related to living conditions and well-being, and to supply data to inform specific policy issues. Classification variables (such as, age, sex, education and income) help delineate population groups for use in the analysis of core data.

3. Content of the 2016 GSS

Canada's rapidly changing demographic profile, along with its accompanying social and economic issues, has led to much discussion concerning the relationship between work, lifestyle and well-being. Gauging the quality of life at work can help diagnose issues relating to productivity, morale, efficiency and equity. Charting patterns of home and leisure activities can reflect the workings of Canadian culture. Bringing these two together will provide insight on the health and well-being of Canadians as they meet the challenges of daily life now and in the future.

The General Social Survey Program's new cycle, Canadians at Work and Home, takes a comprehensive look at the way Canadians live by incorporating the realms of work, home, leisure, and overall well-being into a single survey. Data users have expressed a strong interest in knowing more about the lifestyle behaviour of Canadians that impact their health and well-being both in the workplace and at home. The strength of this survey is its ability to take diverse information Canadians provide on various facets of life and combine them in ways not previously possible with surveys that covered one main topic only.

3.1 Background questions

After the roster is completed and a respondent selected (using the application's automated tool), the date of birth of the respondent and their marital status are recorded. This is followed by establishing the relationship of household members to the respondent, the respondent's place of birth and their immigrant status. In addition, the respondent is asked a series of questions on main activity, work history and various employment characteristics.

3.2 Life at work

This part of the survey paints a picture of how Canadians feel about their workplace by asking questions on how they relate to their work, their teams, their colleagues, supervisors and subordinates. In addition, the survey explores topics such as work ethic, work intensity and distribution, compensation and employment benefits, job satisfaction and meaning, intercultural workplace relations, and discrimination and competition in the workplace. In addition to these variables being of interest on their own, many are correlated with productivity, motivation, absenteeism/tardiness, accidents, mental or physical health and general life satisfaction rendering them essential when exploring both micro and macro issues of individuals and the economy. In addition to covering Canadians with paid jobs, certain sub-groups, such as the self-employed, those looking for work, retirees and students are asked a separate set of questions relevant to their specific situations.

3.3 Life at home

To explore the diversity of today's families and living arrangements, questions include satisfaction with the quantity and quality of time spent together as a family and the division of labour in the household. Other topics under the "Life at home" banner include sports and outdoor activities, cultural participation, the use of technology and nutritional awareness. In an effort to combat an increasingly sedentary lifestyle, rising rates of obesity and their attendant health-related problems, interest in physical activity of all kinds has been intensifying. Responses to the sports and outdoor activities questions will provide some sense of how active Canadians are today. In the area of culture, we note that technology has had an increasingly transforming role by impacting the way we consume, produce and distribute cultural products. The survey explores both traditional forms of participation and expressions of culture through technological devices.

3.4 Work-life balance

A long-time staple of the GSS program, time and stress have been important organizing pillars of different survey cycles. Questions in the 2016 GSS probe the incidence of paid work encroaching on family time, family responsibilities interfering with paid work, and levels of satisfaction with the balance between paid work and home life. Closely related to work-life balance, though not exclusive to it, are questions on the amount and source of stress that Canadians experience in their day-to-day lives.

3.5 Health, well-being and resilience

Self-rated health and subjective well-being are long-standing questions in GSS surveys used over the years to gauge the health and well-being of respondents. While self-rated health has become an important mainstay in social surveys, and tracks very well with other objective measures of health, subjective well-being has traditionally been used as a measure of overall satisfaction with life. The 2016 GSS complements these by a number of new questions that will enable a more detailed look into how Canadians' assess and evaluate their lives. A set of questions on domain satisfaction provides a detailed breakdown of satisfaction across different spheres of life, such as standard of living, health,

achievements, and many others. A module on life opportunities asks about the differences in generational opportunities and aims to illuminate concerns over social mobility, social cohesion and an inclusive society. Questions on outlook ask whether people feel that their overall and financial lives will be easier or harder in the future. Finally, a set of questions on resilience investigates the ability of individuals to bounce back from crises or adversity, the most common definition of resilience.

3.6 Sociodemographic variables

This section provides a variety of socio-demographic measures—many of which are repeated each year in the GSS—concerning respondents, their spouses/partners and parents in order to support the analysis of Canadian families and individuals. This cycle of the GSS includes place of birth, immigration status, aboriginal identity and visible minority status, religion of respondent and its importance, language, as well as sexual orientation.

Questions on personal and household income were removed from the questionnaire and these data are now obtained by merging records from the respondent's fiscal files and the GSS.

4. New content and changes in previously existing content

The 2016 GSS on *Canadians at Work and Home* is the first cycle on this topic. While some modules have been collected in previous surveys, most have not. This section will outline all changes not just in previously existing questions but also in processes related to survey development, collection, processing and methodology.

4.1 Summary of key changes

1) Core content

All new content and revisions to some of the existing questions were tested by Statistic Canada's Questionnaire Design Resource Centre (QDRC) in a two-phase process. In the fall of 2014, the content was tested using a paper questionnaire. Then, in the spring of 2015, the electronic application was tested for usability by respondents. From February to mid-March 2016, a full-scale Pilot Test was conducted both as an electronic self-administered questionnaire (EQ) and as a computer assisted telephone interview (CATI).

2) Relationship of selected respondent

In 2015, the full household composition matrix, previously completed as part of Entry, was replaced by Relationship to Selected Respondent (RSR) to reduce interview time. This one-directional roster was used by the 2016 GSS.

3) Socio-demographic classification

Also in 2015, many survey specific socio-demographic questions were replaced by Statistics Canada's harmonized content questions (i.e., standardized modules for household survey variables, such as education, labour force, aboriginal identity, birth place, citizenship, self-rated health and religion). Harmonized content modules contain standard concepts, definitions, classification and wording for multiple collection modes. This new standardized content is for the most part very similar to the previous concepts used by the GSS, but in some cases required adjustments to the traditionally derived variables.

4) Income

Starting in 2015, personal and household income questions were no longer asked as part of the General Social Survey. Income information was obtained instead through a linkage to tax data for respondents who did not object to this linkage. Respondents were notified of the planned linkage before and during the survey. Those who objected to the linking had their objections recorded. For these individuals, no linkage to their tax data took place. Linking to tax data diminishes respondent burden and increases data quality both in terms of accuracy and in response rates. Moreover, GSS 2016 was the first GSS cycle to use family income (obtained from a linkage to tax data) instead of the previously used household income. (See Section 7.3 for more details.)

5) Frame

The 2016 GSS on Canadians at Work and Home uses the redesigned GSS frame, which integrates data from sources of telephone numbers (landline and cellular) available to Statistics Canada and the Address Register (AR). This new frame includes “cell phone only” households, a growing population not covered by the previous Random Digit Dialling (RDD) frame. The sampling unit is also different and is defined as groups of telephone numbers. (See Section 5.3 for more details.)

6) Coding

The North American Industry Classification System (NAICS) 2012 and National Occupational Classification (NOC) 2016 were used for industry and occupation coding.

7) Processing

Most of the ongoing data processing steps are standard, including consistency edits and family edits. Two aspects of processing, however, are still relatively new and merit a more detailed description.

One is Common Tools, used across the Social, Health and Labour Statistics field at Statistics Canada. From the start of questionnaire development through processing and dissemination, these new common tools are designed to streamline questionnaire specifications and processing steps with the aim of improving efficiency, coherence and consistency across surveys. The majority of new procedures are invisible to users, except for those related to the data dictionary. All surveys processed using common tools have variable names of 8 characters or less and the following reserve codes:

6 Valid skip
7 Don’t know
8 Refused
9 Not stated

In addition to the standard common tools procedures, some work-arounds were created for the 2016 GSS to accommodate the fact that common tools used at that time were not able to handle EQ data. The majority of work-arounds were hand-coded in SAS and ran separately. In the final processing stage, both the EQ and the CATI data were merged into one dataset.

It is very important to note that there is no way to separate “don’t know”, “refused”, and “not stated” in the EQ data. As a result, once the data are merged (which is the case for both the Main file and the PUMF) “don’t know”, “refused”, and “not stated” cannot be considered as separate categories. They must be merged and treated as “not stated.” In contrast, the “valid skip” category stands on its own since it refers to respondents who were not asked a particular question because their characteristics precluded them from being asked.

Tax data linkage: Linkage with tax records, was successful and fiscal information was available for 91.9% of 2016 GSS respondents that did not object to the linkage.

8) Derived variables

In past cycles', some derived variables were created from single variables in the survey.
The survey variables were renamed for the PUMF. In cycle 28, this practice was stopped, meaning that some derived variables found in past cycles have been replaced by the actual variable in the survey. For example, in 2009 DIS_SEX was derived from DIS_Q110. In 2014, DIS_SEX was removed and only the actual question number and acronym in the survey were kept. This change does not impact the amount of information available to the user.

In addition to the standard derived variables, 7 further derived variables were created for this cycle:

1. "WORKSIZE" (Determines whether the respondent's workplace is a small, medium-sized or large business
2. "IMMSTAT" (Determines whether respondent is an immigrant or not)
3. "RECIMM" (Determines whether respondent is a recent immigrant or not, i.e., came to Canada in the last 10 years)
4. "CUREMPLO" (Determines if respondent is currently employed as a paid worker or self-employed)
5. "RETIRED" (Determines if respondent is currently retired or not)
6. "RETYEARS" (Determines the number of years since the most recent retirement)
7. "HARASSED" (Determines if the respondent was ever a victim of bullying (discriminated/treated unfairly) in the workplace

9) Weights

Starting in 2013, the use of a new sampling frame and a new definition of the sampling unit have led to a new weighting strategy for the GSS (see Section 8.1). Additionally, the bootstrap weighting strategy has been changed from mean bootstrap to **standard bootstrap weights** (see Appendix F for more information on how to use standard bootstrap weights).

4.2 Comparability of estimates

While much of the content of this survey is brand new, there are some modules that have appeared in previous GSS cycles (for example, sports, culture, some work-related variables, the life satisfaction question, stress in daily life and some others). For these, it is important to point out that any significant change in survey methodology (as outlined above) can affect the comparability of the data over time. It is impossible to determine with certainty whether, and to what extent, differences in a variable are attributable to an actual change in the population or to changes in the survey methodology.

Consequently, at every stage of processing, verification and dissemination, considerable effort was made to produce data that are precise in their level of detail, and to ensure that the published estimates are of good quality in keeping with Statistics Canada standards.

5. Survey and sample design

Data for 2016 General Social Survey (GSS) on Canadians at Work and Home was collected from August 2 to December 23, 2016. Please see the following sections for descriptions of the target population, stratification, the frame, the sampling strategy, the sample size and sample allocation.

5.1 Target population

The target population for the 2016 GSS included all persons 15 years of age and older in Canada, excluding:

1. Residents of the Yukon, Northwest Territories, and Nunavut; and
2. Full-time residents of institutions.

5.2 Stratification

In order to carry out sampling, each of the ten provinces were divided into strata (i.e., geographic areas). Many of the Census Metropolitan Areas[1] (CMAs) were each considered separate strata. This was the case for St. John's, Halifax, Saint John, Montreal, Quebec City, Toronto, Ottawa, Hamilton, Winnipeg, Regina, Saskatoon, Calgary, Edmonton and Vancouver.

All CMAs not on this list are located in Quebec, Ontario and British Columbia, with the exception of Moncton. Three more strata were formed by grouping the remaining CMAs (except Moncton) in each of Quebec, Ontario and British Columbia. Finally, the non-CMA areas of each of the ten provinces were also grouped to form ten more strata, for a total of 27 strata. Moncton was added to the non-CMA stratum for New Brunswick.

5.3 Frame

The survey frame was created using two different components:

- Lists of telephone numbers in use (both landline and cellular) available to Statistics Canada from various sources (telephone companies, Census of population, etc.);
- The Address Register (AR): List of all dwellings within the ten provinces.

The Address Register (AR) was used to group together all telephone numbers associated with the same valid address. About 86% of available telephone numbers were linked to the AR. The records resulting from this linkage could possess more than one telephone number (grouped by the address). The other 14% of telephone numbers not linked to the AR were also included on the frame[2]. The combination of those two components resulted in the survey frame. The rationale for using all the telephone numbers (linked and not linked to the AR) was to ensure a good coverage of all households with telephone numbers.

When more than one telephone number was attached to a record, they were sorted by source and by type of telephone number (landline telephone numbers first and cellular telephone numbers last). The first telephone number was considered the best telephone number available to reach the household.

Please note that for the remaining sections of this document, the word "record" will refer to the grouping of telephone numbers that consists of our sampling unit on the survey frame.

5.4 Sampling strategy

Each record in the survey frame was assigned to a stratum within its province. A simple random sample without replacement of records was next selected in each stratum.

The frame for GSS was created using several linked sources, such as the Census of population, administrative data files and billing files. Compared to the previous random digit dialing frame, coverage

[1] Based on 2011 Census geography

[2] About 9% of these telephone numbers were grouped using address information from administrative sources. Each of the remaining telephone numbers constitutes a single record on the frame.

was improved (though over coverage and under coverage may still exist). Households without telephones were excluded from the survey population. Survey estimates were adjusted (weighted) to represent all persons in the target population, including those not covered by the survey frame.

For the 2016 GSS, 89.8% of the selected telephone numbers reached eligible households. To be eligible, a household had to include at least one person 15 years of age or older. During collection, households that did not meet the eligibility criteria were terminated after an initial set of questions that determined eligibility.

A respondent was then randomly selected from each household to complete an electronic questionnaire or to respond to a telephone interview.

5.5 Sample size and allocation

The target sample size (i.e. the desired number of respondents) for Cycle 30 was 20,000 while the actual number of respondents was 19,609. For each province, minimum sample sizes were determined that would ensure certain estimates would have acceptable sampling variability at the stratum level. Once these stratum sample size targets had been met, the remaining sample was allocated to the strata in a way that balanced the need for precision of both national-level and stratum-level estimates.

6. Collection and response rate

6.1 Collection

Data for the 2016 GSS was collected electronically as a self-completed questionnaire (EQ) as well as via computer assisted telephone interviews (CATI). Respondents were interviewed in the official language of their choice. Proxy interviews were not permitted.

All interviewing took place using centralized telephone facilities in five of Statistics Canada's regional offices, with calls being made from approximately 9:00 a.m. to 9:30 p.m. Mondays to Fridays. Interviewing was also scheduled from 10:00 a.m. to 5:00 p.m. on Saturdays and 1:00 p.m. to 9:00 p.m. on Sundays. The five regional offices were: Halifax, Sherbrooke, Sturgeon Falls, Winnipeg and Edmonton. Interviewers were trained by Statistics Canada staff in telephone interviewing techniques using CATI, as well as in survey concepts and procedures. The majority of interviewers had experience interviewing for previous GSS cycles.

Interviewers were instructed to make all reasonable attempts to obtain a completed interview with the randomly selected member of the household. Those who at first refused to participate were re-contacted up to two more times to explain the importance of the survey and to encourage their participation. 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.

Interviewer manuals are not included in this documentation package but can be made available by contacting Statistics Canada (see Section 11).

Data for the 2016 GSS were collected from August 2nd to December 23rd 2016. The self-completed electronic questionnaire (EQ) were offered until September 30, at which point the EQ portal closed and all incomplete cases were transferred to be done via CATI. The sample comprised one wave.

The following steps summarize the collection process and the movement of cases from EQ to CATI. Once in CATI, cases could not be transferred back to EQ.

1. Paper introductory letters were sent to households whose telephone numbers were able to be matched to addresses. This comprised approximately 80% of cases. The letter provided a secure access code that a member of the household aged 15 or over was asked to use to log in to the survey.
2. Once in the survey application, this household member, the "rosterer," was asked to confirm that his/her age was 15 years or over, verify that the primary phone number shown on screen belonged to someone in the household, and provide the number of landlines and cell-phones that belonged to the household. If the listed phone number did not belong to anyone in the household the case was immediately transferred to CATI. Similarly, if the listed phone number was a business number but there were additional phone numbers associated with the household, the case was immediately transferred to CATI.
3. Finally the rosterer was asked to list the names of all individuals residing in the household, along with their age and sex.
4. The application then randomly selected someone from the household as the respondent, who will be asked to complete the content part of the survey. If the selected respondent is the rosterer, this person can simply continue with the content part of the survey.
5. If the selected respondent was not the rosterer (this person is referred to as the "secondary respondent"), an email address was requested that could be used to contact the secondary respondent. If an email address was not provided, the case was transferred to CATI and an interviewer would attempt to make contact.
6. With a valid email address, an email invitation is sent with a new secure access code for logging in. In the event that the secondary respondent does not complete the survey online by the deadline, the case is also transferred to CATI.

6.2 Response rate

The overall response rate for the 2016 GSS was 50.8%.

The 2016 sample was selected using the new GSS frame, which necessitated some adjustments in the methodology used to calculate the response rates. Starting in 2013, the new frame includes "cell phone only" households, a population that was not covered under the previous RDD sample frame. Addition of "cell phone only" households to the frame was essential since this population constitutes a constantly growing portion of the population and coverage had been steadily declining with the previous frame. While the addition of these households is necessary for coverage of the Canadian population, this population is harder to reach. Others factors that affect comparability of the response rate over time is the way in which status (in-scope, out-of-scope) is determined under the new design and the use of two different modes (electronically and via telephone) to collect data.

7. Processing

7.1 Data capture

Using CATI, responses to survey questions were entered directly into computers as the interview progressed. The CATI data capture program allowed a valid range of codes for each question, had built in edits, and automatically followed the flow of the questionnaire. The data output was transmitted electronically to Ottawa.

For the self-completed electronic questionnaire, the respondents were capturing their answers directly into the application. Built-in edits and flows were programmed to ensure the CATI and EQ applications were identical. The data output was securely transmitted via Statistics Canada secure portal.

7.2 Coding

Several questions allowed for write-in responses. These responses were coded into existing categories (where a match was possible), grouped into new categories or left in “other-specify” (if a match with an existing category was not possible or the frequencies were too small to create a new category). Where possible (e.g., occupation, industry, language, education, country of birth, religion), coding followed standard classification systems used by the population census and Statistics Canada’s harmonized content program.

7.3 Edit and imputation

All survey records were subjected to computer edits throughout the course of the interview. The CATI system identified “out-of-range” values as they were entered. As a result, the interviewer could immediately resolve such problems with the respondent. If interviewers were unable to correctly resolve the detected errors, they could bypass the edit and forward the data to Head Office for resolution. Interviewer comments were reviewed and taken into account during Head Office editing.

Head Office edits performed the same checks as the CATI system, as well as more detailed edits. Records with missing or incorrect information were, in a small number of cases, completed, corrected deterministically, or imputed from other information on the questionnaire.

Non-response was not permitted for questions required for weighting. Values were imputed in the rare case where the sex of the respondent was missing. The imputation was based on a detailed examination of the data and of other useful variables, such as the age and sex of other household members and the interviewer's comments.

In 2016, personal income questions were not asked as part of the survey. Income information was obtained instead through a linkage to tax data for respondents who did not object to this linkage. Income information was obtained from the 2015 T1FF for 90.6% of the respondents. Missing information for all other respondents was imputed. Contrary to GSS 2015, the **family income** (i.e., linking directly to a variable on the T1FF that corresponds to the census family income) was used for GSS 2016 instead of the household income. In total, a family income value was obtained for 89.9% of households. Missing information for all other respondents was imputed.

7.4 Creation of combined and derived variables

A number of variables on the file were derived from information collected on the questionnaires. In some cases, the derived variables are straightforward and involve collapsing of categories. In other cases, two or more variables were combined to create a new variable. The data dictionary identifies which variables are derived and the nature of their derivation.

8. Estimation

When a probability sample is used, as is the case for the GSS, the principle behind estimation is that each person selected in the sample represents (in addition to himself or herself) several other persons not in the sample. For example, in a simple random sample of 2% from a population size of 1,000, each person in the sample represents 50 persons in the population. The number of persons represented by a given person in the sample is usually known as the weight or weighting factor of the sampled person.

The PUMF contains responses and associated information from 19,609 respondents.

Three weighting factors were placed on the PUMF. They are listed and explained below:

WGHT_PER: This is the basic weighting factor for analysis at the person level, i.e. to calculate estimates of the number of persons (non-institutionalized and aged 15 or over) having one or several given

characteristics. WGHT_PER should be used for all person-level estimates. For example, to estimate the number of persons who reported that their use of technology saved them time, the value of WGHT_PER is summed over all records with this characteristic.

In addition, respondents were split (randomly) into two approximately equal sub-samples. Half of the respondents were asked the questions in the cultural participation module and the other half were asked the questions in the sports activities module.

As a result of splitting the sample, the following sets of weights were created.

WGHT_CSP: This is the weighting factor for analysis at the person level created using the sample of persons in the Cultural Activities module. For example, to estimate the number of persons who went to a cultural or artistic festival in the last 12 months, WGHT_CSP should be summed over all records with this characteristic. This weight is zero for respondents who completed the sports and activities modules.

WGHT_SNT: This is the weighting factor for analysis at the person level created using the sample of persons asked the questions in sports activities module. For example, to estimate the number of persons who regularly participated in sports during the last 12 months, WGHT_SNT should be summed over all records with this characteristic. This weight is zero for respondents who completed the cultural participation module.

In addition to the estimation weights, bootstrap weights have been created for the purpose of design-based variance estimation[3].

8.1 Weighting of persons

As mentioned previously, the records on the survey frame are groups of telephone numbers. A simple random sample of those records was selected in each stratum. Therefore, each record within a stratum has an equal probability of selection.

This probability is equal to:

$$\frac{Number\ of\ records\ sampled\ in\ the\ stratum}{Number\ of\ records\ in\ the\ stratum\ from\ the\ survey\ frame}$$

1) Initial weight calculation

Certain households in the survey frame had a probability of being reached through more than one record. This was possible since groupings of telephone numbers were subject to error.

As mentioned previously, telephone numbers belonging to the same valid address were grouped together on the survey frame. However, for a few cases, the grouping of those telephone numbers might be erroneous (i.e. all the telephone numbers grouped together do not belong to the same household). In addition the remaining telephone numbers that could not be linked to addresses were also included in the frame. It is possible that some of those telephone numbers could reach households already covered by the telephone numbers linked to addresses.

As a result, a series of questions were added to the survey to establish the prevalence of these situations.

[3] Three sets of 500 standard bootstrap weights are available for the 2016 GSS on Canadians at Work and Home: WTBS_001 to WTBS_500 for the 19,609 respondents; WSBS_001 to WSBS_500 for respondents asked the questions in sports activities module; WCBS_001 to WCBS_500 for respondents asked the questions in the cultural participation module.

Several adjustments were made to the initial probability of selection to account for the fact that such households had a higher probability of being selected (i.e. they could be contacted through more than one group of telephone numbers). Therefore, the initial weight is the inverse of this adjusted probability of selection. The resulting initial weight is a household weight.

2) Removal of out-of-scope records

Telephone numbers associated with businesses, institutions or other out-of-scope dwellings, as well as numbers not in service or any other non-working numbers are all examples of out of-scope telephone numbers for this survey. Records with all telephone numbers out-of-scope are simply removed from the process, leaving only in-scope records in the sample. These in-scope records keep the same initial weight as described in the previous step.

3) Three-stage non-response adjustment

Weights for responding telephone numbers were adjusted to represent non-responding telephone numbers.

Non-responding telephone numbers were grouped into three types: those with some auxiliary information available (in particular, a complete roster of household members), those with auxiliary information from various sources available to Statistics Canada and those with no auxiliary information.

This non-response adjustment was done in three stages. In the first stage, adjustments were made for complete non-response (i.e., households for which no auxiliary information was available). This was done independently within each stratum. In the second stage, adjustments were made for non-response with auxiliary information from sources available to Statistics Canada. These households had some auxiliary information which was used to model propensity to respond. In the third stage, adjustments were made for partial non-response. These households had some auxiliary information which was used to model propensity to respond. The last two adjustments were done independently within each region[4]. The combination of these three adjustments is referred to as Factor 1.

Non-responding telephone numbers were then dropped.

4) Person weight calculation

A person weight was calculated for the respondent by multiplying the household weight by the number of persons 15 years of age or older in the household.

This step produces a person weight, which can be calculated as:

Initial Household Weight x Factor 1 x Number of Eligible Household Members.

5) Adjustment of person weights to external totals

The person weights were adjusted several times using a raking ratio procedure. This procedure ensures that, based on the survey's total sample, estimates are produced that match certain external reference totals. Two sets of external references were used for this survey, both of them population totals: for stratum (geographic), and for age-sex groups by province.

It should be noted that persons living in households without telephone service (or telephone service not covered by the frame) are included in the external references even though such persons were not sampled.

5a) Stratum Adjustment

[4] The five geographical regions are: Atlantic, Quebec, Ontario, Prairies and British Columbia.

An adjustment was made to the person weights for records within each stratum (geographic) in order to make population estimates consistent with the corresponding projected population counts. This was done by multiplying the person weight for each record within the stratum by the following ratio:

$$\frac{Projected\ Population\ Count\ for\ the\ Stratum}{Sum\ of\ the\ Person\ Weights\ for\ the\ Stratum}$$

5b) Province - age - sex adjustment

The next weighting step adjusts the weights so they agree with projected province-age-sex population distributions. Projected population counts were obtained for males and females within the following sixteen age groups:

15-19	20-24	25-29	30-34
35-39	40-44	45-49	50-54
55-59	60-64	65-69	70-74
75-79	80-84	85-89	90 +

For each of the resulting classifications, the person weights for records within the classification were adjusted by multiplying by the following ratio:

$$\frac{Projected\ Province - Age - Sex\ Group\ Population\ Count}{Sum\ of\ the\ Province - Age - Sex\ Group\ Person\ Weights}$$

When sample sizes were small, adjacent age group data for the same province and sex were combined before this adjustment was made.

5c) Raking Ratio Adjustments

The weights of each respondent were adjusted several times using a raking ratio procedure. This procedure ensured that estimates produced for stratum and province-age-sex totals would agree with the external reference totals. This adjustment was made by repeating steps 5a) and 5b) of the weighting procedures until each repetition of the step made a minimal adjustment to the weights.

6) Final person weight

The weight produced at the end of step 5) is the final person weight (WGHT_PER) placed on the Main File.

7) Person weight based on the split sample

To take into account the split sample design of the questionnaire, two additional person weights WGHT_CSP and WGHT_SNT were created using the same approach as the person weight. WGHT_CSP is zero for respondents who were not asked questions from the cultural participation module and WGHT_SNT is zero for respondents who were not asked questions from the sports activities module.

8.2 Weighting policy

Users are cautioned against releasing unweighted tables or performing any analysis based on unweighted survey results. As was discussed in Section 8.1, several weight adjustments were performed

that depended on the province, stratum, age and sex of the respondent. Sampling rates as well as non-response rates varied significantly from province to province, and non-response rates varied with demographic characteristics. For example, non-respondents are often more likely to be males and more likely to be younger. In the responding sample, 2.0% were males between the ages of 20 and 24, while in the overall population, approximately 4.1% were males between 20 and 24. Therefore, it is clear that unweighted sample counts cannot be considered representative of the survey target population.

The total number of households in the survey's scope was estimated at 38,569*.* Among these households, 19,609 usable responses were obtained, which gives a response rate of 50.8%. The distribution of the non-response and response categories is given in the table below:

Source	Number	%
1. Household non-response	14,482	37.5
2. Refusal (person level)	1,165	3.0
3. Other non-response (person level)	3,313	8.6
4. Response (CATI)	13,520	35.1
5. Response (EQ)	6,089	15.8
Total Households	**38,569**	**100.0**

In all, the number of non-response cases is estimated to 18,960 cases. Categories 2 and 3 show non-response occurring after the respondent was selected. The "other non-response' category (3) includes cases where no response could be obtained because of language difficulties or other problems. Responses obtained from Electronic Questionnaire (EQ) represents 31.1% of the 19,609 responses obtained.

8.3 Types of estimates

Two types of "simple" estimates are possible from the results of the General Social Survey. These are qualitative estimates (estimates of counts or proportions of people possessing certain qualities or characteristics) and quantitative estimates involving quantities of averages. More complex estimation and analyses are covered in Section 8.5.

8.3.1 Qualitative estimates

The target population for the GSS was non-institutionalized persons aged 15 and older, living in the ten provinces. Qualitative estimates are estimates of the number or proportion of this target population possessing certain characteristics. The number of people (4,657,522) who describe their state of health as excellent (SRH_110 = 1) is an example of this kind of estimate. These estimates are readily obtained by summing the person weights (WGHT_PER) for the records possessing the characteristic of interest. This estimate does not, however, adjust for non-response to the question in any way.

If we make the assumption that those who either refused to answer the question or who responded 'Don't Know' have the same distribution as those who responded, then an adjusted estimate can be made. To do this, the proportion of the target population with this characteristic is estimated by excluding respondents with a 'Not Stated' or 'Don't Know' answer to question SRH_110 and calculating the ratio of the total of the weights of those respondents who answered that their state of health was 'Excellent' (SRH_110=1) to that of all respondents who answered the question (SRH_110=1, 2, 3, 4, or 5). This proportion is then multiplied by the size of the target population to produce the final estimate (it should be noted that this adjustment does not have to be done, but it can be if needed):

$$4{,}680{,}203 \;=\; 30{,}078{,}789 \;\text{x}\; \frac{4{,}657{,}522}{29{,}933{,}021}$$

30,078,789 is the estimated number of persons aged 15 and over in the population (target population). 29,933,021 is the sum of the weights of all respondents who answered question SRH_110 (i.e. SRH_110 = 1,2,3,4 or 5). When the proportion of responses that are 'Don't Know' or 'Refused' are high, the differences between the two estimates will be large.

8.3.2 Quantitative estimates

Some variables on the General Social Survey PUMF are quantitative in nature (e.g. age, number of weeks worked in the past 12 months). From these variables, it is possible to obtain estimates such as the average number of weeks worked in the past 12 months. These quantitative estimates are of the following ratio form:

$$\text{Estimate (average)} = \frac{X}{Y}$$

The numerator (X) is a quantitative estimate of the total for the variable of interest (for example, the number of weeks worked in the past 12 months) for a given sub-population (for example, males who worked in the past 12 months). In this example, X would be calculated by multiplying the person weight (WGHT_PER) by the variable of interest (WET_110) when it is known, 1 <= WET_110 <= 52, (i.e. not equal to '96', '97', '98' or '99'), and summing this product over all records for males who worked i.e. SEX=1 and (1 <= WET_110 <= 52), which yields 482,951,168.

The denominator (Y) is the qualitative estimate of the number of persons within that sub-population (males who worked in the past 12 months). In this example, Y would be calculated by summing the person weight (WGHT_PER) over all male respondents with 1 <= WET_110 <= 52, yielding 10,975,964.

The two estimates X and Y are derived independently and then divided to provide the quantitative estimate. The average number of weeks is then calculated to be:

$$\frac{482{,}951{,}168}{10{,}975{,}964} = 44.0$$

8.4 Guidelines for analysis

As detailed in Section 5 of this document, the respondents from the GSS do not form a simple random sample of the target population. Instead, the survey had a complex design, with stratification, multiple stages of selection and unequal selection probabilities for respondents. Using data from such complex surveys presents analytical challenges because the survey design and selection probabilities affect the estimation and variance calculations that should be used.

The GSS used a stratified design, with significant differences in sampling fractions between strata. Thus, some areas were over-represented in the sample (relative to their populations) while some other areas were relatively under-represented; this means that the unweighted sample was not representative of the target population, even if there was no non-response. Non-response rates may vary by demographic group, making the unweighted sample even less representative.

The survey weights must be used when producing estimates or performing analyses in order to account as much as possible for the geographic over- and under-representation and for the under- or over-representation of age-sex groups or months of the year in the unweighted file. While many analysis

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

For many analysis techniques (for example linear regression, logistic regression, estimation of rates and proportions and analysis of variance), a method exists which can make the variances calculated by the standard packages more meaningful. If the weights for the data, or for the subset of the data that is of interest, are rescaled so that the average weight is one (1), then the variances produced by the standard packages will be more reasonable; they still will not take into account the stratification and clustering of the sample's design, but they will take into account the unequal probabilities of selection. This rescaling can be accomplished by dividing each weight by the overall average weight before the analysis is conducted. Section 9 describes sampling variability and data reliability in more detail.

8.5 Estimating number of persons by using WGHT_PER on the main file

As previously mentioned, a basic person weight has been assigned to each sampled individual and, as described in Section 7.1, these weights have been adjusted to reflect the age and sex composition of the various provincial populations as estimated by Statistics Canada for each month covered by Cycle 30

$$\sum_{i=1}^{n=19,609} WGHT_{PER} = 30{,}078{,}789^{1}$$

[1] Estimate of the number of persons aged 15 and over in the population

In general, when an estimate is based on the person as the unit of observation, the Main File and WGHT_PER should be used. Example of this are the average number of weeks worked by persons aged 25 to 29 years old, the percentage of persons whose main activity in the past 12 months was going to school, and the number of people aged between 25 and 44 who are currently attending school, college, CEGEP or university.

The last example would be calculated as follows: WGHT_PER would be summed up for all records on the main file with 2<= AGEGR10 <= 3 and ESC1_01 = 1, giving an estimate of 954,136 persons aged 25 to 44 who are currently attending school, college, CEGEP or university.

9. Release guidelines and data reliability

It is important for users to become familiar with the contents of this section before publishing or otherwise releasing any estimates derived from the General Social Survey PUMF.

This section of the documentation provides guidelines to be followed by users. With the aid of these guidelines, users of the PUMF should be able to produce figures consistent with those produced by Statistics Canada and in conformance with the established guidelines for rounding and release. The guidelines include four broad sections: Minimum Sample Sizes for Estimates; Sampling Variability Policy; Sampling Variability Estimation; and Rounding Policy.

9.1 Minimum sample size for estimates

Users should determine the number of records on the PUMF which contribute to the calculation of a given estimate. This number should be at least 15 in the case of persons or households. When the number of contributors to the weighted estimate is less than 15, the weighted estimate should generally not be

released regardless of the value of the coefficient of variation. If it is, it should be with great caution and the insufficient number of contributors associated with the estimate should be prominently noted.

9.2 Sampling variability guidelines

The estimates derived from this survey are based on a sample of persons. Somewhat different figures might have been obtained if a complete census had been taken using the same questionnaire, interviewers, supervisors, processing methods, etc. as those actually used. The difference between the estimates obtained from the sample and the results from a complete count taken under similar conditions is called the sampling error of the estimate.

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

Over a large number of observations, randomly occurring errors will have little effect on estimates derived from the survey. However, errors occurring systematically will contribute to biases in the survey estimates. Considerable time and effort was made to reduce non-sampling errors in the survey. Quality assurance measures were used at each step of the data collection and processing cycle to monitor the quality of the data. These measures included the use of highly skilled interviewers, extensive training of interviewers with respect to the survey procedures and questionnaire, observation of interviewers to detect problems of questionnaire design or misunderstanding of instructions, and coding and edit quality checks to verify the processing logic.

9.2.1 Non-sampling errors

A major source of non-sampling errors in surveys is the effect of non-response on the survey results. The extent of non-response varies from partial non-response (failure to answer one or a few questions) to total non-response. Total non-response occurred because either the interviewer was unable to contact the respondent, no member of the household was able to provide the information (perhaps due to a language problem), or the respondent refused to participate in the survey. Total non-response was handled by adjusting the weight of households who responded to the survey to compensate for those who did not respond.

In most cases, partial non-response to the survey occurred when the respondent did not understand or misinterpreted a question, refused to answer a question, or could not recall the requested information.

9.2.2 Sampling errors

Since it is an unavoidable fact that estimates from a sample survey are subject to sampling error, sound statistical practice calls for researchers to provide users with some indication of the magnitude of this sampling error.

Although the exact sampling error of the estimate, as defined above, cannot be measured from sample results alone, it is possible to estimate a statistical measure of sampling error, the standard error, from the sample data. Using the standard error, confidence intervals for estimates (ignoring the effects of non-sampling error) may be obtained under the assumption that the estimates are normally distributed about the true population value. 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 virtually certain that the differences would be less than three standard errors.

Since the absolute size of the sampling error of an estimate is often less important than its relative size (relative to the estimate itself) the standard error is not always the best measure of sampling error. For example, a standard error of 10 for an estimate of 20 would generally be taken as indicating that the estimate is a poor one, while the same standard error for an estimate of 1,000 would generally indicate a good estimate. For this reason the size of the sampling error is often expressed relative to the size of the estimate, as the coefficient of variation (c.v.). The coefficient of variation of an estimate is obtained by dividing the standard error of the estimate by the estimate itself, and the resulting fraction is usually expressed as a percentage. In the above example, the first estimate has a c.v. of 50% (10/20), while the second has a c.v. of 1% (10/1,000).

The choice between using the standard error or the CV as a measure of sampling variability is one the user should make based on his/her specific analysis. Guidelines for publishing estimates using the CV are given in the next section.

With enough observations, the user can proceed to calculating variances and coefficients of variation using the bootstrap weights provided with the data (see Section 9.2.3 for guidelines to follow when using coefficients of variation and Section 9.3 for more details on the appropriate software to use for bootstrap weights).

9.2.3 Guidelines for release of estimates

When considering releasing *and*/*or* publishing an estimate from the PUMF, users should consult the table below and follow the guideline that matches the coefficient of variation of the estimate.

Type of Estimate	**Coefficient of Variation**	**Policy Statement**
1.With Moderate Sampling Variability	0.0% to 16.5%	Estimates can be considered for general unrestricted release. No special notation is required.
2. With High Sampling Variability	16.6% to 33.3%	Estimates can be considered for general unrestricted release but should be accompanied by a warning cautioning users of the high sampling variability associated with the estimates.
3. With Very High Sampling Variability	33.4% or over	Estimates should generally not be released, but when they are it should be with great caution and the very high sampling variability associated with the estimate should be prominently noted.

9.3 Variance estimation using bootstrap weights

In order to determine the quality of the estimate and to calculate the CV, the standard deviation must be calculated. Confidence intervals also require the standard deviation of the estimate. The GSS uses a multi-stage survey design and calibration, which means that there is no simple formula that can be used to calculate variance estimates. Therefore, an approximate method was needed. The bootstrap method is used because the sample design and calibration needs to be taken into account when calculating

variance estimates. With the use of available software to compute variances and the help of bootstrap weights (discussed in the next subsection), the method is fairly easy for users.

This technique involves dividing the records on the microdata file into subgroups (or replicates) and determining the variation in the estimates from replicate to replicate. The replicates are formed by selecting, independently within each stratum, a simple random sample with replacement of n-1 of the n units in the sample. Note that since the selection is with replacement, a unit may be chosen more than once. A bootstrap weight based on the bootstrap sample is calculated for each sample unit in the stratum. This process (selecting simple random samples, recalculating weights for each stratum) is repeated B times, where B is large, yielding B different initial bootstrap weights. The GSS typically uses B=500, to produce 500 bootstrap weights.

These weights are then adjusted according to the same weighting process as the regular person weights: non-response adjustment, calibration and so on. The end result is 500 final bootstrap weights for each unit in the sample. The variation among the 500 possible estimates based on the 500 bootstrap weights is related to the variance of the estimator based on the regular weights and can be used to estimate it.

9.4 Rounding

In order for estimates produced from the GSS microdata files to correspond to those produced by Statistics Canada, users are urged to adhere to the following guidelines regarding the rounding of such estimates. It may be misleading to release unrounded estimates, as they imply greater precision than actually exists.

9.4.1 Rounding guidelines

1) Estimates of totals in the main body of a statistical table should be rounded to the nearest thousand using the normal rounding technique (see definition in Section 9.4.2).

2) 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 thousand units using normal rounding.

3) Averages, proportions, rates and percentages are to be computed from unrounded components and then are to be rounded themselves to one decimal using normal rounding.

4) Sums and differences of aggregates and ratios are to be derived from corresponding unrounded components and then rounded to the nearest thousand units or the nearest one decimal using normal rounding.

5) In instances where, due to technical or other limitations, a different rounding technique is used, resulting in estimates different from Statistics Canada estimates, users are encouraged to note the reason for such differences in the released document.

9.4.2 Normal rounding

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, the number 8499 rounded to thousands would be 8000 and the number 8500 rounded to thousands would be 9000.

10. Additional information

For additional information please contact:

Survey Manager
Anna Kemeny
Social and Aboriginal Statistics Division
(613) 854-8420
Anna.kemeny@canada.ca

Data from the survey is available through published reports, special request tabulations, and the microdata file. The microdata file will be available from the Social and Aboriginal Statistics Division of Statistics Canada. Tabulations can be obtained at a cost that will reflect the resources required to produce the tabulation.

To receive a copy of the microdata file or to order special tabulations, please contact:

Client Services
Social and Aboriginal Statistics Division
Statistics Canada
Sasdclientservices-dsseaserviceaclientele.statcan@canada.ca

Appendix A – Citizenship codes

CNTRYTEXT	CTZCODE
Other - Specify	001
Angola	024
United States (USA)	103
Belize	201
Costa Rica	202
El Salvador	203
Guatemala	204
Honduras	205
Mexico	206
Nicaragua	207
Antigua and Barbuda	302
Bahamas	304
Barbados	305
Cuba	308
Dominica	309
Dominican Republic	310
Grenada	311
Haiti	313
Jamaica	314
Saint Kitts and Nevis	319
Saint Lucia	320
Saint Vincent and the Grenadines	321

Trinidad and Tobago	322
Argentina	401
Bolivia	402
Brazil	403
Chile	404
Colombia	405
Ecuador	406
Guyana	409
Paraguay	410
Peru	411
Suriname	412
Uruguay	413
Venezuela	414
Austria	501
Belgium	502
France and dependencies	503
Germany	505
Liechtenstein	506
Luxembourg	507
Monaco	508
Netherlands and dependencies	509
Switzerland	511
Bulgaria	521
Czechoslovakia	522
Czech Republic	523
Estonia	524
Hungary	525
Latvia	526
Lithuania	527
Romania	529
Slovakia	530
Union of Soviet Socialist Republic (USSR)	531
Belarus	533
Moldova, Republic of	534
Russian Federation	535
Ukraine	536
Ireland, Republic of	541
Ireland (Eire)	542
British	543
United Kingdom - British citizens	544
Denmark	546

Finland	547
Iceland	548
Norway	549
Sweden	550
Guernsey	553
Jersey	555
Albania	561
Andorra	562
Bosnia and Herzegovina	563
Croatia	564
Greece	566
Italy	567
Macedonia, Republic of	568
Malta	569
Portugal	571
San Marino	572
Yugoslavia	573
Slovenia	574
Spain	575
Holy See (Vatican City State)	576
Serbia and Montenegro	578
Kosovo	583
Panama	591
Benin	601
Burkina Faso	602
Cape Verde	603
Côte d'Ivoire (Ivory Coast)	604
Gambia	605
Ghana	606
Guinea	607
Guinea-Bissau	608
Liberia	609
Mali	610
Mauritania	611
Niger	612
Nigeria	613
Senegal	615
Sierra Leone	616
Togo	617
Burundi	621
Comoros	622

Djibouti	623
Eritrea	624
Ethiopia	625
Kenya	626
Madagascar	627
Malawi	628
Mauritius	629
Mozambique	631
Rwanda	633
Seychelles	634
Somalia	635
Tanzania, United Republic of	636
Uganda	637
Zambia	638
Zimbabwe	639
Algeria	651
Egypt	652
Libya	653
Morocco	654
Sudan	655
Tunisia	656
Western Sahara	657
Cameroon	662
Central African Republic	663
Chad	664
Congo, Republic of	665
Equatorial Guinea	666
Gabon	667
Sao Tome and Principe	668
Congo, Democratic Republic of (Zaire)	669
Botswana	681
Lesotho	682
Namibia	683
South Africa, Republic of	684
Swaziland	685
Afghanistan	701
Cyprus	702
Iran	703
Turkey	704
Bahrain	705
Iraq	706

Israel	707
Jordan	708
Kuwait	709
Lebanon	710
Oman	711
Palestine	712
Qatar	713
Saudi Arabia	714
Syria	715
United Arab Emirates	716
West Bank and Gaza Strip (Palestine)	717
Yemen	718
Armenia	720
Azerbaijan	721
Georgia, Republic of	722
Kazakhstan	723
Kyrgyzstan	724
Tajikistan	725
Turkmenistan	726
Uzbekistan	727
China, People's Republic of	732
Japan	734
Korea, North	735
Korea, South	736
Mongolia	739
Taiwan	740
Brunei Darussalam	751
Cambodia	752
Indonesia	753
Laos	754
Malaysia	755
Burma (Myanmar)	756
Myanmar (Burma)	756
Philippines	757
Singapore	758
Thailand	759
Viet Nam	760
Bangladesh	771
Bhutan	772
India	773
Maldives	774

Nepal	775
Pakistan	776
Sri Lanka	777
East Timor (Timor Leste)	781
Kurdistan	782
Gaza Strip	784
Australia	802
Fiji	804
Kiribati	807
Marshall Islands	808
Micronesia, Federated States of	809
Nauru	810
New Zealand	812
Palau	813
Papua New Guinea	814
Solomon Islands	816
Tonga	817
Tuvalu	818
Vanuatu	820
Samoa	822
Poland	825
Porto Rico	831
United States Minor Outlying Islands	832
Stateless	906
United Kingdom - dependent territories	908
Canada	991
111 - No more countries	995

Appendix B – Country of birth codes

CNTRYTEXT	CNTRYCODE
Canada	11124
Greenland	11304
Saint Pierre and Miquelon	11666
United States (USA)	11840
Belize	12084
Costa Rica	12188
El Salvador	12222
Guatemala	12320
Honduras	12340
Mexico	12484
Nicaragua	12558
Panama	12591
Netherlands Antilles	13000
Antigua and Barbuda	13028
Bahamas	13044
Barbados	13052
Bermuda	13060
Virgin Islands, British	13092
Cayman Islands	13136
Cuba	13192
Dominica	13212
Dominican Republic	13214
Grenada	13308
Guadeloupe	13312
Haiti	13332
Jamaica	13388
Martinique	13474
Montserrat	13500
Curaçao	13531
Aruba	13533
Saint-Maarten (Dutch part)	13534
Bonaire, Saint Eustatius and Saba	13535
Puerto Rico	13630
Saint Barthélemy	13652
Saint Kitts and Nevis	13659
Anguilla	13660
Saint Lucia	13662
Saint Maarten (French part)	13663

Saint Vincent and the Grenadines	13670
Trinidad and Tobago	13780
Turks and Caicos Islands	13796
Virgin Islands, United States	13850
Argentina	14032
Bolivia	14068
Brazil	14076
Chile	14152
Colombia	14170
Ecuador	14218
Falkland Islands (Malvinas)	14238
South Georgia and the South Sandwich Islands	14239
French Guiana	14254
Guyana	14328
Paraguay	14600
Peru	14604
Suriname	14740
Uruguay	14858
Venezuela	14862
Austria	21040
Belgium	21056
France	21250
Germany	21276
Liechtenstein	21438
Luxembourg	21442
Monaco	21492
Netherlands	21528
Switzerland	21756
Austria-Hungary	22000
Czechoslovakia	22010
Union of Soviet Socialist Republic	22020
Bulgaria	22100
Belarus	22112
Czech Republic	22203
Estonia	22233
Hungary	22348
Latvia	22428
Lithuania	22440
Moldova	22498
Poland	22616
Romania	22642

Russian Federation	22643
Slovakia	22703
Ukraine	22804
Denmark	23208
Faroe Islands	23234
Finland	23246
Åland Islands	23248
Iceland	23352
Ireland	23372
Ireland, Republic of	23373
Norway	23578
Svalbard and Jan Mayen	23744
Sweden	23752
Great Britain	23825
England	23826
Northern Ireland	23827
Scotland	23828
United Kingdom	23829
Wales	23830
Guernsey	23831
Jersey	23832
Isle of Man	23833
Yugoslavia	24000
Albania	24008
Andorra	24020
Bosnia and Herzegovina	24070
Croatia	24191
Gibraltar	24292
Greece	24300
Holy See (Vatican City State)	24336
Italy	24380
Malta	24470
Montenegro	24499
Portugal	24620
San Marino	24674
Serbia	24688
Slovenia	24705
Spain	24724
Macedonia (Region)	24807
Macedonia, Republic of	24808
Kosovo	24983

Cape Verde	31132
Benin	31204
Gambia	31270
Ghana	31288
Guinea	31324
Ivory Coast	31384
Liberia	31430
Mali	31466
Mauritania	31478
Niger	31562
Nigeria	31566
Guinea-Bissau	31624
Saint Helena and Ascension	31654
Senegal	31686
Sierra Leone	31694
Togo	31768
Burkina Faso	31854
Burundi	32108
Comoros	32174
Mayotte	32175
Ethiopia	32231
Eritrea	32232
Djibouti	32262
Kenya	32404
Madagascar	32450
Malawi	32454
Mauritius	32480
Mozambique	32508
Réunion	32638
Rwanda	32646
Seychelles	32690
Somalia	32706
Zimbabwe	32716
Uganda	32800
Tanzania	32834
Zambia	32894
Algeria	33012
Libya	33434
Morocco	33504
Sudan	33736
Tunisia	33788

Egypt	33818
Angola	34024
Cameroon	34120
Central African Republic	34140
Chad	34148
Congo, Republic of	34178
Congo, Democratic Republic of (Zaire)	34180
Congo, Unspecified	34181
Equatorial Guinea	34226
Gabon	34266
Sao Tome and Principe	34678
Botswana	35072
Lesotho	35426
Namibia	35516
South Africa, Republic of	35710
Swaziland	35748
Kurdistan	41000
Afghanistan	41004
Azerbaijan	41031
Bahrain	41048
Armenia	41051
Cyprus	41196
Georgia	41268
Georgia, Republic of	41268
Gaza Strip	41275
Palestine	41276
West Bank and Gaza Strip (Palestine)	41277
Iran	41364
Iraq	41368
Israel	41376
Kazakhstan	41398
Jordan	41400
Kuwait	41414
Kyrgyzstan	41417
Lebanon	41422
Oman	41512
Qatar	41634
Saudi Arabia	41682
Syria	41760
Tajikistan	41762
United Arab Emirates	41784

Turkey	41792
Turkmenistan	41795
Uzbekistan	41860
Yemen	41887
Korea	42000
China	42156
Taiwan	42158
Hong Kong Special Administrative Region	42344
Japan	42392
Korea, North	42408
Korea, South	42410
Macao Special Administrative Region	42446
Mongolia	42496
Brunei Darussalam	43096
Burma (Myanmar)	43104
Cambodia	43116
Indonesia	43360
Laos	43418
Malaysia	43458
Philippines	43608
East Timor (Timor Leste)	43626
Singapore	43702
Viet Nam	43704
Thailand	43764
Bangladesh	44050
Bhutan	44064
British Indian Ocean Territory	44086
Sri Lanka	44144
India	44356
Maldives	44462
Nepal	44524
Pakistan	44586
Oceania	50000
American Samoa	51016
Australia	51036
Solomon Islands	51090
Christmas Island	51162
Cocos (Keeling) Islands	51166
Cook Islands	51184
Fiji	51242
French Polynesia	51258

Kiribati	51296
Guam	51316
Nauru	51520
New Caledonia	51540
Vanuatu	51548
New Zealand	51554
Niue	51570
Norfolk Island	51574
Northern Mariana Islands	51580
United States Minor Outlying Islands	51581
Micronesia, Federated States of	51583
Marshall Islands	51584
Palau	51585
Papua New Guinea	51598
Pitcairn	51612
Tokelau	51772
Tonga	51776
Tuvalu	51798
Wallis and Futuna	51876
Samoa	51882
Antarctica Adjacent Islands	60000
Antarctica	61010
Bouvet Island	61074
French Southern Territories	61260
Heard Island and McDonald Islands	61334
Born at Sea	80000
Other - Specify	90000

Appendix C – Religion codes

Religion	Code
Aboriginal spirituality	**201000**
Agnostic	301000
Ahmadiyya	105051
Amish	102270
Anabaptist	102271
Ancestor Worship	201101
Anglican	102020
Anglo-Catholic	102021
Animism	**201100**
Antiochian Orthodox Christian	103010
Apostolic	102381
Apostolic Christian Church	102030
Armenian Apostolic	103021
Armenian Catholic	101011
Armenian Orthodox	**103020**
Associated Gospel	102040
Atheist	302000
Baha'i	105010
Baptist	102050
Blackfoot	201001
Born-again Christian	102131
Brethren in Christ	102060
Buddhist	105020
Bulgarian Orthodox	103051
Canadian and American Reformed Churches	102071
Catholic	101000
Chaldean Catholic	101012
Charismatic Renewal	102080
Christadelphian	102090
Christian	102130
Christian and Missionary Alliance	102100
Christian Congregation	102120
Christian Fundamentalist	102384
Christian or Plymouth Brethren	102420
Christian Orthodox	103052
Christian Reformed Church	102140
Church of God	102150
Church of Jesus Christ of Latter-day Saints	102160

Church of Scotland/Scottish Presbyterian	102431
Church of the Nazarene	102170
Community of Christ	102481
Coptic Orthodox	103030
Covenant Church	102432
Cree	201002
Déné	201003
Doukhobor	102037
Druid	206001
Druze	105052
Dutch Reformed Church	102190
Eastern Catholic	101013
Eastern religions	**105000**
Eastern Rite Catholic	101014
Eckankar	105001
Ethiopian Orthodox	103053
Evangelical	102383
Evangelical Baptist	102051
Evangelical Covenant Church	102382
Evangelical Free Church	102211
Evangelical Missionary Church	102210
Four Square Gospel Church	102401
Fourth Way	209001
Free Church	102434
Free Methodist	102220
Free Reformed Church	102070
Free Thinker	303000
Full Gospel	102402
German Lutheran	102261
Glad Tidings	102403
Gnostic	210000
Gospel	102404
Gospel Hall	102406
Grace Communion International	102551
Greek or Byzantine Catholic	101015
Greek Orthodox	103040
Hindu	105040
Humanist	304000
Hutterite	102230
Iglesia ni Cristo	102385
Interdenominational Christian	102240

Ismali	105053
Jainism	105060
Jedi	211001
Jehovah's Witness	102250
Jewish	**104000**
Jewish Orthodox	104001
Laestasdian Lutheran	102262
Longhouse	201004
Lutheran	102260
Macedonian Orthodox	103054
Manitou	201005
Mar Thoma Syrian Church/Marthomite	102386
Maronite	101016
Melkite	101017
Mennonite	102272
Mennonite Brethen	102273
Messianic Jew/Jewish Christian	102387
Methodist	102280
Metropolitan Community Church	102388
Midewin	201006
Mission de l'Esprit Saint	102310
Moravian Church	102330
Mormon	**102480**
Muslim	**105050**
Native American Church	102389
New Age	204000
New Apostolic	102340
Non-denominational Christian	102360
Orthodox	103000
Other - Specify	900000
Other Catholic	101010
Other Christian	**102380**
Other Orthodox	103055
Pagan	**206000**
Pentecostal	**102400**
Polish National Catholic Church	101020
Presbyterian	**102430**
Protestant	102000
Quaker	102450
Raelian	211002
Rastafari	207000

Reformed	102436
Reformed Presbyterian	102435
Reinlander Mennonite	102274
Revival Centre	102391
Roman Catholic	101030
Romanian Orthodox	103060
Russian Orthodox	103070
Salvation Army	102490
Satanism	208000
Science of Mind/Religious Science	201102
Scientology	**209000**
Serbian Orthodox	103080
Seventh-day Adventist	102010
Shaker	102392
Shamanism	206002
Shi'a	105054
Shinto	105080
Sikh	105090
Sommerfeld Mennonite	102275
Southern Baptist	102052
Spiritual Baptist	102053
Spiritualist	102500
Standard Church	**102510**
Sunni	105055
Swedenborgian (New Church)	102350
Syrian Catholic	101018
Taoist	105100
The Christian Church (Disciples of Christ)	102180
Ukrainian Catholic	101040
Ukrainian Orthodox	103090
Unitarian	102520
United Church	102530
Unity - New Thought - Pantheist	102511
Vineyard Christian Fellowship	102393
Wesleyan	102540
Wicca	206003
Worldwide Church of God	**102550**
Zoroastrian	105071

Appendix D – Sports codes

Sport	Code
Adaptive or parasport	**1000**
Adventure racing	**1005**
Archery	**1010**
Arctic sports	**1015**
Badminton	**1020**
Ball Hockey	**1025**
Baseball	**1030**
Basketball	**1035**
BMX	**1040**
Bobsleigh	**1045**
Bowling, five pin	**1050**
Bowling, ten pin	**1055**
Boxing	**1060**
Broomball	**1065**
Canoe-Kayak	**1070**
Cheerleading	**1075**
Combined Track and Field (Triathlon, pentathlon, decathlon, etc.)	**1080**
Combined Winter Sports (Biathlon, Nordic combined, Alpine Combined)	**1085**
Competitive weightlifting	**1090**
Cricket	**1095**
Cross country skiing	**1100**
Curling	**1105**
Cycling	**1110**
Diving	**1115**
Downhill/Alpine Skiing	**1120**
Equestrian	**1125**
Fencing	**1130**
Field Hockey	**1135**
Figure Skating	**1140**
Football	**1145**
Freestyle Skiing	**1150**
Golf	**1155**
Gymnastics	**1160**
Handball (team, 4 walls)	**1165**
Ice hockey	**1170**
In-line hockey	**1175**
In-line skating	**1180**
Kickboxing and Mixed Martial Arts	**1185**
Lacrosse	**1190**
Lawn bowling	**1195**

Luge	**1200**
Marathon	**1205**
Martial Arts (Karate, Jujitsu, Judo, Aikido, Tae Kwon Do, Kung Fu, Capoeira, etc.)	**1210**
Mountain Biking	**1215**
Mountain Boarding	**1220**
Netball	**1225**
Race walking	**1230**
Racquetball	**1235**
Ringette	**1240**
Rowing	**1245**
Rugby	**1250**
Running - road	**1255**
Sailing/Yachting	**1260**
Shooting	**1265**
Skateboarding	**1270**
Skeleton	**1275**
Ski jumping	**1280**
Snowboarding	**1285**
Snowshoeing	**1290**
Soccer	**1295**
Softball	**1300**
Speed Skating	**1305**
Squash	**1310**
Swimming	**1315**
Synchronized Swimming	**1320**
Table Tennis/ping pong	**1325**
Tennis	**1330**
Track and field	**1335**
Ultimate Frisbee	**1340**
Volleyball	**1345**
Water Polo	**1350**
Water skiing/Wakeboarding	**1355**
Windsurfing/Sailboarding	**1360**
Wrestling	**1365**

Appendix E – Major field of study codes

Major field of study codes	Code
No postsecondary certificate, diploma or degree	**01**
Education	**02**
Visual and performing arts, and communications technologies	**03**
Humanities	**04**
Social and behavioural sciences and law	**05**
Business, management and public administration	**06**
Physical and life sciences and technologies	**07**
Mathematics, computer and information sciences	**08**
Architecture, engineering, and related technologies	**09**
Agriculture, natural resources and conservation	**10**
Health and related fields	**11**
Personal, protective and transportation services	**12**
Other	**13**

Appendix F – Tips for using GSS standard bootstrap weights

A survey weight variable with a corresponding set of 500 standard bootstrap weight[5] variables are provided with many GSS microdata files in order that a full design-based approach may be taken for doing analysis with the data.

A design-based approach to analysis first involves using the survey weight variable for obtaining weighted estimates of the quantities of interest. Then, additional information about the survey design is used in order to make estimates of the variances[6] (and covariances) of these estimated quantities. In the case of many GSS microdata files, this additional information is in the form of 500 survey bootstrap weight variables. The design-based estimates and variance estimates can then be used for making the inferences required in the analysis.

The form of a bootstrap variance estimate can be described briefly as follows:

Let $\hat{\beta}$ be the weighted estimate of quantity of interest, β, computed using the survey weight variable w, and let $\hat{\beta}^{(b)}$ be an estimate obtained in exactly the same manner, except for substituting the bth bootstrap weight variable $w^{(b)}$ for the survey weight variable w, b=1,2,…500. This yields the bootstrap estimates $\hat{\beta}^{(1)},\ldots,\hat{\beta}^{(500)}$ of β. Then the usual bootstrap estimate of the variance of $\hat{\beta}$ is

$$\hat{V}_B(\hat{\beta}) = \frac{1}{500}\sum_{b=1}^{500}\left(\hat{\beta}^{(b)} - \hat{\beta}\right)^2 . \qquad (1)$$

[5] Unlike previous years, GSS now uses standard bootstrap weights. Special attention should be given to formula (1) as it is different from the formula for the mean bootstrap weights.

[6] The variance that is estimated in a design-based approach is the variability in an estimate due to re-sampling by exactly the same design from the same finite population.

If $\hat{\beta}$ is a vector instead of a single value, such as if $\hat{\beta}$ is the set of coefficients of a model, then the matrix of estimates of the variances and covariances of the elements of $\hat{\beta}$ is $\hat{V}_B(\hat{\beta}) = \frac{1}{500}\sum_{b=1}^{500}\left(\hat{\beta}^{(b)} - \hat{\beta}\right)\left(\hat{\beta}^{(b)} - \hat{\beta}\right)'$. (The value "500" in the formula is due to the fact that we have 500 different series of bootstrap weights. If the number of bootstrap samples should change from 500, then the values in formula (1) would need to change.)

Survey bootstrapping is just one replication approach that may be used in order to obtain design-based variance estimates with survey data. While several commercial software packages for design-based analysis offer replication approaches for variance estimation, they usually do not specify bootstrapping as one of these approaches. However, due to the similarity in the form of the variance estimate for the bootstrap and for the particular replication method called BRR, programs that can carry out variance estimation by this latter approach with user-supplied replication weights can be used to obtain bootstrap variance estimates[7]. In particular, in these software, the 500 bootstrap weights provided in the GSS microdata files need to be designated as 500 BRR weights.

In the sections below, instructions will be given for implementing bootstrap variance estimation with GSS microdata, using 3 different commercial software packages that can carry out some design-based analysis for BRR: Stata 9 or 10, SUDAAN and WesVar. In all GSS cycles where bootstrap weights are provided, the names given to these bootstrap variables in the user documentation are wtbs_001 to wtbs_500[8]. The name of the survey weight variable is usually wght_per.

Stata 9 or 10

Beginning with Version 9, the commercial software package Stata added some replication approaches for carrying out design-based variance estimation in its survey analysis commands. One replication approach offered is the BRR approach, and it is this approach that would be specified when analyzing GSS data.
In order to specify this approach, the following is recommended:

1. Before using any of the survey analysis commands, use a "svyset" statement to declare the data to be survey data, to designate the variables that contain information about the survey design and to specify the method for variance estimation. Settings made by "svyset" are saved with a dataset when (or if) a dataset is saved. The form of the svyset statement to be used with a GSS analysis dataset would have the following form:

svyset [pweight=wght_per], vce(brr) fay(0) brrweight(wtbs_001-wtbs_500) mse

Declaring **pweight=wght_per** tells Stata that the survey weight (which is often called the probability weight) is the variable wght_per.

The option **vce(brr)** states that the variance estimation approach to use is BRR.

The option **brrweight(wtbs_001-wtbs_500)** states that the names of the BRR weight variables are **wtbs_001, wtbs_002, …, wtbs_500.** This option can also be designated as **brrweight(wtbs_*)** provided there are no variables other than the bootstrap weight variables whose names begin with "wtbs_".

[7] For a more detailed description see Phillips (2004)
[8] Please note that in previous GSS cycles (Cycle 26 and earlier), the variables wtbs_001 to wtbs_500 were mean bootstrap weights. Beginning with cycle 27 of GSS (2013), the variables wtbs_001 to wtbs_500 are standard bootstrap weights.

The option **fay(0)** states that the BRR variance estimation approach used does not require a Fay's adjustment. A Fay's adjustment was required when using mean bootstrap weights but, starting with cycle 27 of GSS, we now use standard bootstrap weights and this adjustment is unnecessary.

Finally, the **mse** option tells Stata to calculate the variance using squared differences between bootstrap estimates and the full-sample estimate of the quantities of interest, as shown in equation (1). If this option is not included, Stata uses squared differences between each bootstrap estimate and the mean of all the bootstrap estimates. Both approaches should yield approximately the same result.

2. There is an extensive list of survey analysis commands in Stata, which take a design-based approach in their computations. These commands, described in the Stata documentation, are implemented through the use of the "svy" prefix along with the names of other estimators. For example, **svy: mean** is the command for estimating population and subpopulation means and estimates of variability taking a design-based approach. When the **svyset** statement precedes all survey commands, the survey commands do not have to contain any information about the design-based approach to be taken. It should be noted that, even though most of the commands that allow the "svy" prefix are also the names of commands for non-survey data, what is estimated, what options are available and what can be done through post-estimation change when the "svy" prefix is added.

SUDAAN

SUDAAN is a commercial software package developed by the Research Triangle Institute specifically for analysis of data from complex sample surveys and other observational and experimental studies involving cluster-correlated data. The SAS-callable version of the software is particularly useful to people familiar with SAS.

Specification of the variance estimation approach to be used by SUDAAN is done in the procedure statement for a particular procedure. Additional sample design statements provide further information required by the program. In particular, to carry out bootstrapping with GSS data, the following is required:

- specify **DESIGN=BRR** in the procedure statement
- include the following WEIGHT statement to identify the survey weight variable:

WEIGHT wght_per;

- include the REPWGT statement to indicate the names of the bootstrap variables on your data file. In particular, for GSS microdata files, this REPWGT statement would have the form:

REPWGT wtbs_001-wtbs_500;

WesVar

WesVar is a software package produced by Westat which carries out various analyses of survey data using exclusively replication methods for variance estimation. One of the methods offered is BRR. Quoting heavily from Phillips (2004), in WesVar, the variance estimation method is specified when creating a new WesVar data file. The resulting file is then used to define workbooks where table and regression requests are carried out. To define a WesVar data file with bootstrap weights:

- Move the replicate weight variables (i.e., wtbs-001 to wtbs_500) to the *Replicates* box..
- Move the survey weight variable (i.e., wght_per) to the *Full sample* box.
- For the mean bootstrap, specify the *Method* as BRR.
- Move analysis variables to the *Variables* box, a unique identifier to the ID box (optional), and save the file.

References

Phillips, Owen (2004) “Using Bootstrap Weights with WesVar and SUDAAN”. The Research Data Centres Information and Technical Bulletin. (Fall) 1(2):1-10. Statistics Canada Catalogue no. 12-002-XIE. http://www.statcan.ca/bsolc/english/bsolc?catno=12-002-X2004002703