

THE 1999 GENERAL SOCIAL SURVEY - CYCLE 13
VICTIMIZATION

Public Use Microdata File Documentation

and

User's Guide

Fall 2000

Aussi disponible en français

The General Social Survey - Cycle 13

Public Use Microdata File Documentation

and User's Guide

Table of Contents

1. Introduction.....	1
2. Objectives of the GSS.....	1
3. Content and Special Features of Cycle 13.....	2
4. Survey and Sample design.....	6
5. Collection.....	8
6. Processing.....	8
7. Estimation.....	12
8. Release Guidelines and Data Reliability.....	29
9. Additional Information.....	35
Appendix A. Approximate Variance Tables	
Appendix B. GSS Cycle 13 Victimization Questionnaire	
Appendix C. GSS Cycle 13 on Victimization: Comparison with Cycle 3, Cycle 8 and the Violence Against Women Survey	
Appendix D. Topical Index to Variables for Main File	
Appendix E. Data Dictionary for Main File	
Appendix F. Record Layout for Main File	
Appendix G. Topical Index to Variables for Incident File	
Appendix H. Data Dictionary for Incident File	
Appendix I. Record Layout for Incident File	

1. INTRODUCTION

This package is designed to enable interested users to access and manipulate the microdata file for the thirteenth cycle of the General Social Survey (GSS), conducted from February through December, 1999, in collaboration with the Canadian Centre for Justice Statistics of Statistics Canada. It contains information on the objectives, methodology and estimation procedures as well as guidelines for releasing estimates based on the survey. It also gives a description of how to correctly use the microdata files.

Appendix A contains the Approximate Variance Tables with examples of their use. Appendix B includes the questionnaire and Appendix C a document comparing Cycle 3, Cycle 8 and the Violence against Women Survey (VAWS) with Cycle 13. Appendix D contains the topical index for the main file variables, the data dictionary for the main microdata file is found in Appendix E and Appendix F contains the main file record layout. Appendix G contains the topical index for the incident file variables, the data dictionary for the incident microdata file is found in Appendix H and Appendix I contains the incident file record layout.

This package is available in machine readable form.

2. OBJECTIVES OF THE GENERAL SOCIAL SURVEY

The GSS program, established in 1985, conducts telephone surveys across the 10 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 emerging issues.

The two primary objectives of the General Social Survey (GSS) 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 immediate information on specific social policy issues of current or emerging interest.

To meet the stated objectives, the data collected by the GSS are made up of three components: Classification, Core and Focus. Classification content consists of variables which provide the means of delineating population groups and for use in the analysis of Core and Focus data. Examples of classification variables are age, sex, education, and income. Core content, such as victimization, is designed to obtain information that monitors social trends or measures changes in society related to living conditions or well-being. Focus content is aimed at the second survey objective of the GSS. This component obtains information on specific policy issues that are of particular interest to certain federal departments or other user groups.

Cycle 13 of the GSS is the third cycle (after Cycles 3 & 8) to collect information on the nature and extent of criminal victimization in Canada. The two previous cycles included questions on

unintentional injuries, i.e. accidents. Since the National Health Population Survey collects information on accidents, these questions have been dropped from Cycle 13. Focus content for Cycle 13 addresses two areas of emerging interest. For the 1999 GSS on victimization, the Interdepartmental Working Group on Family Violence has sponsored modules on spousal and senior abuse and the department of the Solicitor General Canada has funded questions measuring public perception toward alternatives to imprisonment.

3. CONTENT AND SPECIAL FEATURES OF THE GSS CYCLE 13

The content of the questionnaire is outlined below.

Control Form

This survey uses the Random Digit Dialing approach to selecting households. The Control Form gathers information about the household composition, including age, sex, and marital status of members. A respondent is randomly selected from among the eligible persons (people 15 years of age or older) in the household. The relationship of all household members to one another is determined and the main questionnaire is completed for the respondent. Proxy interviews are not accepted. Also included in this section is a series of telephone questions used for weighting that ask respondents about the number of telephone numbers in their home and how many are for business, computer or fax use only.

Section A: Perceptions, History and Risk

This section introduces respondents to the survey and the issues to follow. This section is designed to measure the extent to which people worry about their personal safety in everyday situations, the extent to which fear imposes limits on their opportunities and freedom of movement, and how they manage threats to their safety in their daily lives.

Specifically, in this section information is collected on a variety of issues dealing with perceptions of crime and the justice system, contact and satisfaction with various aspects of the criminal justice system, crime prevention measures, frequency of evening activities, and fear of crime. Through funding from the department of the Solicitor General Canada, a new set of questions on public perceptions toward alternatives to imprisonment has been added. In addition, questions related to public perceptions of the prison and parole systems have been included to complement the questions dealing with public perceptions of the police and courts.

One open-ended question asks the respondent whether there is anything else they do to increase their personal safety that has not been mentioned. This question is the first opportunity for the respondents to speak in their own words and serves to build rapport with the interviewer.

Section B: Criminal Victimization Screening Section

This section collects information on the type of crimes inflicted on the respondent over the past 12 months, as well as the number of occurrences of each type. Each time a crime is reported in this section a Crime Incident Report (Section V) is completed. Questions in this section are, for the most part, a repeat of the 1993 GSS questions. Again respondents are asked to include incidents committed by family and non-family. Due to the addition of the spousal and senior abuse questions (Sections C through K), respondents are asked to exclude physical and sexual assaults committed by current and previous spouses or common-law partners. In addition, respondents 65 years of age and older are asked to exclude physical assaults by children and caregivers and sexual assaults by caregivers.

Section C and E: Emotional and Financial Abuse by Current and Previous Spouse/ Partner

Questions related to controlling and emotionally abusive behaviour on the part of a marital partner have been taken from the 1993 Violence Against Women Survey. The intent of these questions is to test theories about links between spousal power and control and spousal violence. In addition, these measures help us better understand the dynamics of an abusive and violent relationship by providing context to reports of violence by spouses. There is substantial evidence to suggest that emotional abuse can be as devastating as actual physical assault. Two additional items have been added to the original scale used in the Violence Against Women Survey. These include threatening to harm or harming someone close and deliberately damaging or destroying property.

Sections D and F: Violence in Current and Previous Marriage/Common-law Relationships

Measures of violence by current or previous marital partners are obtained through a number of categories of violent acts ranging from threats of violence to threats or use of guns or other weapons, and sexual assault. The scale of items used is taken from the Violence Against Women Survey. Research suggests that questions itemizing violent incidents into discrete categories of behaviour are necessary in order to counteract denial and unwillingness to identify experiences as assault or violence.

The method in which the scale is applied, however, differs from the approach used in the Violence Against Women Survey, in that respondents are asked all of the items in the scale. In addition, while the first two items are always presented as the first two items in the scale, the remaining eight questions are randomized. In the Violence Against Women Survey respondents who did not disclose incidents of violence skipped out of the sequence after every three items with the question “Has he ever been violent toward you in any other way?”

Emphasis in this section is on violence within the relationship and not on quantifying individual acts or events. Respondents are asked to specify the number of different occasions that their spouse has been violent toward them. In the case of marriages of less than 5 years, respondents are asked

whether the violence occurred before they were married or living together, during the marriage or during a temporary separation. In the case of previous partners, respondents are asked whether the violence happened after separation, and if so, whether it increased after separation. Respondents who report one or more acts of violence by a spouse or partner are asked to respond to questions in an Abuse Report (Sections L and M).

Section G and J: Emotional and Financial Abuse of Seniors by Children and Caregivers

Similar to the questions related to controlling and emotionally abusive behaviour on the part of a marital partner, these questions also build on the 1993 Violence Against Women Survey. Two questions are used to screen respondents 65 years of age and older into the section related to children, including whether they have ever had any children and how frequently they have had contact with their children. In addition, in this section the sex of the child who has been emotionally or financially abusive is determined. In the case of senior abuse by caregivers, one screening question is used concerning the use of paid or unpaid caregivers.

While items concerning sexual jealousy and the insistence on knowing the respondent's whereabouts at all times are not included in these sections, three items have been added. These items are related to financial abuse. In the case of caregivers, respondents are not asked about attempts to limit their contacts with family or friends, nor are they asked about control over family finances.

Sections H and K: Violence against Seniors by Children and Caregivers

These sections use the same approach and scale of questions that is used in the sections concerning spousal violence. The item on sexual violence, however, is not asked of respondents in the case of children. Respondents who report one or more acts of violence by a child or caregiver are asked to respond to questions in an Abuse Report (Section N and P).

Section K ends with a question about lifetime experiences of victimization. This question is asked of all survey respondents. The intent of this question is to look at lifetime experiences of victimization and to offer respondents an opportunity to indicate incidents of crime that happened outside the 12 month and 5 year timeframe.

Section Q: Classification

Section Q provides background characteristics of respondents regardless of whether they have been a victim of crime. This section contains a variety of socio-demographic and health measures that contribute to the analysis of risk factors and other correlates of people's fear and victimization.

Most classification questions have evolved with each cycle of the GSS. For example, questions concerning mother and father's place of birth were not asked in Cycle 8, but are now asked in each GSS cycle. This cycle also includes questions related to mother and father's education. Questions

about current spouse's use of alcohol, their education, and whether their current spouse has been out of work and looking for work are also included, as these are important for identifying correlates of spousal violence and high risk offenders.

Section V: Crime Incident Report

A Crime Incident Report is completed for every crime reported in Section B of the Victimization Questionnaire. The Crime Incident Report collects information on characteristics of victimization incidents (e.g. month, place), physical, financial and emotional consequences of victimization, offender characteristics (e.g. age, gender), whether medical attention was sought, whether the police were contacted and the level of satisfaction with the police.

A large part of the 1999 Crime Incident Report is similar to the 1993 Crime Incident Report. However a few changes have been made. Inquiries as to whether the incident was related to the use of alcohol or drugs by the perpetrator or the victim are asked of all victims; no longer are victims of sexual assault excluded. Furthermore, all victims are asked about attempting to or obtaining civil or criminal compensation; no longer are victims of violent crimes excluded. In addition, while the question concerning the time of the incident has been dropped, some questions have been added. These include: two questions concerning hate crime, questions related to the people to whom the respondent may have spoken about the incident or the services he or she may have contacted or used, two questions on the age and sex of the perpetrator(s), one question about victim/offender mediation, and a question to assess how the respondent was affected by the incident. Finally, the categories of possible actions taken by the police have been expanded, and one item has been added to the list of reasons why a respondent may not have contacted the police.

Section L, M, N and P: Abuse Reports

The Abuse Reports collect detailed information about violence in current or previous marriages/common-law relationships, and violence against seniors by children or caregivers. Each report contains questions about: the impact of the experience for the victim (physical injury and emotional trauma); whether anyone else was threatened or harmed in the incident(s); use of criminal compensation; who they turned to for help; involvement and satisfaction with the police; actions taken by the police; reasons for contacting or not contacting the police; and interest in participating in victim/offender mediation programs.

In the spousal violence Abuse Report, respondents are also asked about the impact that police intervention has had on their partner's behaviour, if they ever feared that their life was in danger, and whether their children ever witnessed the violence.

An open-ended question allowing victims of spousal or senior abuse to speak in their own words asks if there is any advice they would give another person in a similar situation.

4. SURVEY AND SAMPLE DESIGN

Data for Cycle 13 of the GSS were collected monthly from February 1999 to December 1999 inclusively. With the exception of extra Newfoundland units purchased late in the year by an external client, the sample was more or less evenly distributed over the 11 months to counterbalance as much as possible the seasonal variation in the information gathered. All of the sample was selected using the Elimination of Non-Working Banks technique. A description of this method is provided in Section 4.3. The target population is discussed in Section 4.1 and the stratification used in the survey design is outlined in Section 4.2.

4.1 Target Population

The target population for the GSS was all persons 15 years of age and over in Canada, excluding:

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

The survey employed Random Digit Dialling (RDD), a telephone sampling method. Households without telephones were therefore excluded. There is evidence, however, that persons living in such households represent less than 2% of the target population. Survey estimates have been adjusted (weighted) to represent all persons in the target population, including those without telephones. The tacit assumption is that, given the small number of people without telephones, their characteristics are not different enough from those of the rest of the target population to have an impact on the estimates. Since no one without a telephone is in the sample, this assumption cannot be verified using GSS data. The characteristics of the population without telephones has been examined using data from the Survey of Consumer Finance and the Household Facilities and Equipment Survey. Telephone ownership is high among virtually all socio-economic groups, but is lowest among the 3% of the population with the lowest household income (less than \$10,000). The telephone ownership rate was 92.6% for this population, while it was over 96% for all other groups.

4.2 Stratification

In order to carry out sampling, each of the ten provinces was divided into strata, i.e. geographic areas.

Many of the Census Metropolitan Areas (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, Vancouver and Victoria. CMAs not on this list are located in Quebec and Ontario. Two more strata were formed by grouping the remaining CMAs in each of these two provinces. Finally, the non-CMA areas of each of the ten provinces were also grouped to form ten more strata. This resulted in 27 strata in all.

The size of the sample for Cycle 13 was 25,876, an increase by a factor of about 2½ from previous

GSS cycles. In addition to having a larger number of strata as described above, allocation to these strata was changed. For each province, minimum sample sizes were determined that would ensure that certain provincial victimization estimates would have acceptably low sampling variability, based on Cycle 8 results. Once these provincial sample size targets had been met, most of the remaining sample was allocated to the provinces in a way that optimized the precision of national-level estimates.

The portion of the sample that had already been allocated to the provinces was assigned to the sub-provincial strata proportionally to their size. Finally, the remainder of the sample was allocated to strata with a high proportion of people in a visible minority (such as Toronto), in order to improve estimates for these groups. An external client purchased extra sample for Newfoundland from September to December.

4.3 Elimination of Non-working Banks RDD Design

The Elimination of Non-Working Banks (ENWB) sampling technique is a method of Random Digit Dialling in which an attempt is made to identify all working banks¹ for an area (i.e., to identify all banks containing at least one number that belongs to a household). Thus, all telephone numbers within non-working banks are eliminated from the sampling frame.

For each province, lists of telephone numbers in use were purchased from the telephone companies and lists of working banks were extracted. Each bank was assigned to a stratum within its province.

A special situation existed in Ontario and Quebec because some small areas are serviced by independent telephone companies for which we did not have lists. Area code / prefixes² from Ontario and Quebec and not on our list files were identified. All banks within these area code / prefixes were generated and added to the sampling frame.

In each stratum, a simple random sample without replacement of telephone numbers was selected by choosing a simple random sample with replacement of banks from the frame, and then randomly generating the last two digits for each bank to obtain the telephone number. The entire monthly sample of telephone numbers was produced before the first day of interviewing for the month. Therefore, a prediction of the percentage of numbers dialled that would reach a household had to be made (this is known as the "hit rate"). The hit rates for the Cycle 12 RDD sample were used to

¹ A bank of telephone numbers is a set of 100 numbers with the same first eight digits (i.e. the same Area Code-Prefix-Bank ID). Thus 613-951-9180 and 613-951-9192 are in the same bank, but 613-951-9280 is in a different bank.

² An area code / prefix is determined by the first six digits of a telephone number, for instance 613-951.

estimate the hit rates for Cycle 13 during the early months of collection. Predicted hit rates were adjusted in later months to reflect the hit rates observed during the previous months of Cycle 13 collection.

For Cycle 13 of the GSS, 45.7% of the numbers dialled reached households. An attempt was made to conduct a GSS interview with one randomly selected person from each household.

5. COLLECTION

Computer assisted telephone interviewing (CATI) was used to collect data for the GSS, using Random Digit Dialling methods. Respondents were interviewed in the official language of their choice. Interviews by proxy were not allowed. Data collection began in February 1999 and continued through to December 1999. The sample was evenly distributed over the 11 months, except for the Newfoundland units purchased during the latter part of the year by an external client. The overall response rate during collection for Cycle 13 was 81.3%.

All interviewing took place using centralized telephone facilities in four of Statistics Canada's regional offices, with calls being made from approximately 09:00 until 21:00, Monday to Saturday inclusive. The four regional offices were: Halifax, Montreal, Winnipeg and Vancouver.

Interviewers were trained by Statistics Canada staff in telephone interviewing techniques using CATI, survey concepts and procedures. Because of the sensitive nature of the focus content, interviewers were also provided with personal preparedness training by a psychologist. The majority of interviewers had previous experience interviewing for the GSS.

It would be too lengthy to include all the survey manuals as part of this documentation package. However, more information can be obtained from Statistics Canada (see Section 9). Shown below is a list of the manuals used in the survey:

GSS Cycle 13 Victimization Training Guide
GSS Cycle 13 Victimization Interviewer's Manual

6. PROCESSING

6.1 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 and built-in edits, and automatically followed the flow of the questionnaire. The information output by the CATI system was transmitted electronically to Ottawa.

6.2 Coding

Several questions allowing write-in responses had this information coded into either new unique categories, or to a listed category if the write-in information duplicated a listed category. Where possible (e.g., occupation, industry, language, education, country of birth, religion), the coding followed the standard classification systems as used in the Census of Population.

6.3 Edit and Imputation

All survey records were subjected to computer edits throughout the course of the interview. The CATI system principally edited flow of the questionnaire and identified out of range values. As a result, such problems could be immediately resolved with the respondent. If the interviewer was unable to correctly resolve the detected errors, it was possible for the interviewer to bypass the edit and forward the data to head office for resolution. All interviewer comments were reviewed and taken into account in head office editing.

Head office edits performed the same checks as the CATI system as well as more detailed edits. Due to the nature of the survey, imputation was not appropriate for most items. Records with missing or incorrect information were, in a small number of cases, corrected or obtained deterministically from other information on the questionnaire. For example, a total of 150 incident reports out of a total of 10,087 were corrected deterministically based on the information provided in Section B of the questionnaire.

The flow editing carried out by head office followed a 'top down' strategy, in that whether or not a given question was considered "on path" was based on the response codes to the previous questions. If the response codes to the previous questions indicated that the current question was "on path", the responses, if any, to the current question were retained, though "don't know" was recoded as 8 (98 or 998, etc.) and refusals were recoded as "Not Stated", i.e. 9 (99 or 999, etc.); if, however, a response was missing to the current question, it was coded as "Not Stated", i.e. 9 (99 or 999, etc.). If the response codes to the previous questions indicated that the current question was "off path" because the respondent was clearly identified as belonging to a sub-population for which the current question was inappropriate or not of interest, the current question was coded as "Not Applicable", i.e. 7 (97 or 997, etc.). However, if the respondent refused or answered "don't know" to a previous branching question, the current off-path question was assigned code 6 (96 or 996, etc.) "Not stated – path unknown". For example, if the respondent answered "Don't know" to Question Q5 "In what country were you born?" the questionnaire flow would have sent the respondent to Question Q8. Consequently, questions Q6A to Q7 would have missing data. These missing items were assigned code 6 "Not stated – path unknown". If however, the respondent answered "Canada" to Question Q5, and data was missing for Question Q6A "In which province or territory", Q6A was assigned code 9 "Not stated", since it was known that the respondent was on path for question Q6A.

The imputation described below was necessary for a small number of cases (56 of 489 reporting

spousal violence in the past 12 months). In order to obtain estimates of the total number of incidents of physical assault and the total number of incidents of sexual assault that occurred in the past 12 months, incidents resulting from spousal violence must be included. Some respondents reported both physical and sexual assault by a spouse in the past 5 years, and some form of violence in the past 12 months. However, it was unknown whether the violence was physical or sexual in nature for any given incident having occurred in the past 12 months. For these incidents, the nature of the violence was imputed as physical assault and the incidents were included in the variable PR_304 rather than PR_101 (see Section 6.4 Creation of Combined and Derived Variables). Consequently, the total number of incidents of physical assault in the past 12 months is slightly overestimated, while the total number of incidents of sexual assault is slightly underestimated. 136 incidents of spousal violence (out of a total of 891 assigned to PR_101 and PR_304) were included in PR_304 through this imputation.

Non-response was not permitted for those items required for weighting. Values were imputed in the rare cases where either of the following were missing: sex or number of residential telephone lines. The imputation was based on a detailed examination of the data and the consideration of any useful data such as the ages and sexes of other household members, and the interviewer's comments. The number of residential telephone lines was assumed to be one (1) when the respondent failed to provide the information.

6.4 Creation of Combined and Derived Variables

A number of variables on the file have been 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 have been combined to create a new variable. The data dictionary identifies which variables are derived and the nature of their derivation.

It should be noted that in 1999, the definition of assault was adjusted to make it more consistent with Criminal Code definitions. In previous victimization cycles, any threat of physical harm was counted as an assault. In Cycle 13, only threats that were face-to-face and that the respondent believed would be carried out were counted as assaults. Analysts are cautioned against making comparisons between measures of victimization from Cycle 13 and Cycle 8 without making the necessary adjustments to Cycle 8. See Appendix C “GSS Cycle 13 on Victimization: Comparison with Cycle 3, Cycle 8 and the Violence Against Women Survey”.

6.4.1 Scenario Variables:

In Section A of the questionnaire, the 32 questions A26A through to A33AP in fact consist of 8 sets of 4 questions each. The sets are known as scenarios and each respondent was asked the questions from only one scenario. In other words, the sample was divided into 8 sub-samples and each scenario was assigned to a different sub-sample. There are only 4 distinct series of questions among the 8 scenarios - each of the 4 series appears in 2 scenarios but with a different order of questions. This was done to help compensate for any bias a given order of questions might create.

The questions deal with the respondent's attitudes towards sentencing of adult and young offenders in hypothetical situations as follows (for precise details of the hypothetical crimes, see questionnaire):

Question Series A (scenarios 1 & 2): crime is a break and enter and it's a first offence

Question Series B (scenarios 3 & 4): crime is an assault with minor injuries and it's a first offence

Question Series C (scenarios 5 & 6): crime is a break and enter and it's a second offence

Question Series D (scenarios 7 & 8): crime is an assault with minor injuries and it's a second offence.

The variable QUSERIES identifies the question series that was asked of the respondent.

Scenarios 1, 3, 5 and 7 ask the questions about adult offenders first and the other scenarios ask the questions about young offenders first.

The answers for all the scenario questions are contained in variables SCEN_A, SCEN_Y, SCEN_AP and SCEN_YP. SCEN_A contains the answers to questions A26A, A27A, ... , A33A with the scenario determining which of these questions was asked of the respondent for a given case. SCEN_Y, SCEN_AP and SCEN_YP correspond analogously to the Y, AP and YP groups of questions, respectively.

Since any given scenario question always appears in two scenarios, analysts interested in the answers to these questions must work with the question series mentioned above (variable QUSERIES), for which the scenarios with identical questions have been combined. Therefore:

SCEN_A contains the answer to whether the respondent feels a prison or non-prison sentence is appropriate for an adult committing the crime specified in QUSERIES as described above.

SCEN_AP contains the answer to whether the respondent feels a year of probation and 200 hours of community work is acceptable as an alternative to prison for an adult committing the crime specified in QUSERIES as described above.

SCEN_Y contains the answer to whether the respondent feels a prison or non-prison sentence is appropriate for a youth committing the crime specified in QUSERIES as described above.

SCEN_YP contains the answer to whether the respondent feels a year of probation and 200 hours of community work is acceptable as an alternative to prison for a youth committing the crime specified in QUSERIES as described above.

6.4.2 Variables with Incidents of Spousal Violence or of Violence against Seniors:

Estimates of total numbers of violent incidents in the past 12 months should include those occurring within violent spousal relationships and violent relationships involving a senior with a

child or caregiver. Violence in these relationships, though, was not examined incident by incident as were other violent victimization incidents, i.e. there are no incident reports for this kind of violence. The number of violent incidents in the past 12 months was recorded, however.

PR_101 contains the number of incidents of sexual assault the respondent experienced from a spouse or ex-spouse within the past 12 months. The number of incidents inflicted by each of the two types of perpetrator (spouse and ex-spouse) is capped at 3, so that 6 is the highest possible value for the variable.

PR_304 contains the number of incidents of physical assault the respondent experienced from a spouse or ex-spouse within the past 12 months. The number of incidents inflicted by each of the two types of perpetrator (spouse and ex-spouse) is capped at 3, so that 6 is the highest possible value for the variable.

SN_101 contains the number of incidents of sexual assault that a senior respondent experienced from a caregiver within the past 12 months. It is capped at 3.

SN_304 contains the number of incidents of physical assault that a senior respondent experienced from a child or caregiver within the past 12 months. The number of incidents inflicted by each of the two types of perpetrator (child and caregiver) is capped at 3, so that 6 is the highest possible value for the variable.

As mentioned in Section 6.3, it was not known whether or not certain incidents were of a sexual nature. To be conservative, these were imputed to be incidents of physical assault.

6.5 Amount of Detail on Microdata File

In order to guard against disclosure, the amount of detail included on this file is less than is available on the master file retained by Statistics Canada. Variables with extreme values have been capped and information for some variables has been aggregated into broader classes (e.g., occupation, religion).

The measures taken to cap, group or collapse data have been indicated in the data dictionary. Variables with a very limited number of observations or referring to small population areas have been excluded from the file. For example, Sections N and P have not been included on the file for these reasons.

7. ESTIMATION

When a probability sample is used, as was the case for the GSS, the principle behind estimation is that each person selected in the sample represents (in addition to himself/herself) several other persons not in the sample. For example, in a simple random sample of 2% of the population, 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.

There are two microdata files from which GSS Cycle 13 estimates can be made. The **Main File** contains questionnaire responses and associated information from 25,876 respondents. Characteristics on this file concern the person as opposed to information about any individual victimization incidents which he or she may have experienced.

Four weighting factors were placed on the Main File. 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 **with the exception of the questions found in section A (the scenarios) that are related to the respondent's feelings on sentencing in various hypothetical crime situations**. For example, to estimate the number of persons who are very worried while waiting for or using public transportation, the value of WGHT_PER is summed over all records with this characteristic.

WGHT_SCE: This weighting factor, as opposed to WGHT_PER, should be used for the questions found in section A (the scenarios) that are related to the respondent's feelings on sentencing in various hypothetical crime situations. For example, to estimate the number of persons who feel that a prison sentence is appropriate for a first-time adult offender found guilty of breaking into a house while the owners are on vacation and taking goods worth \$400, the value of WGHT_SCE is summed over all records with this characteristic. In the appropriate context, both WGHT_PER and WGHT_SCE give the number of persons in the population that the record represents. The reason that WGHT_PER cannot be used in the scenario context is that only one quarter of the sample was asked any given scenario question. The other three quarters were asked similar but not identical questions. For scenario questions a given record therefore represents about four times as many persons in the population as for other characteristics and a separate weighting factor is required. More information and examples on how to use the scenarios and the WGHT_SCE weighting factor can be found in Section 7.6.

WGHT_HSD: This weighting factor can be used to estimate the number of households with a given characteristic. For example, to estimate the number of households that live in low-rise apartments, WGHT_HSD should be summed over all records with this characteristic.

WGHT_ABU: This weighting factor is required to estimate the number of victimization incidents that occurred over the past 12 months within certain violent relationships, namely those with spousal or ex-spousal violence, or with violence against seniors involving a son, daughter or caregiver. It should therefore only be used for estimates involving variables PR_101, PR_304, SN_101 and SN_304 (see Section 6.4.2). In these cases, reports were completed for the violent relationship as a whole rather than for individual victimization incidents (see Incident File below).

The number of individual incidents involved over the past 12 months was recorded, however, and their nature, i.e. physical or sexual assault, imputed (see Section 6.3). Suppose one wants to estimate the number of victimization incidents of a certain kind that occurred within these types of violent relationships over the past 12 months. The number of incidents of the kind in question should be multiplied by WGHT_ABU and summed over all records. For example, to estimate the number of physical assault incidents within spousal or ex-spousal relationships over the past 12 months, PR_304 should be multiplied by WGHT_ABU and summed over all records.

The second microdata file is the **Incident File**. The records on this file contain reports of victimization incidents. Each victimization incident experienced by a respondent of the survey is included on one of the file's records, unless the incident occurred within any of the following violent relationships: those with spousal or ex-spousal violence, or with violence against seniors involving a son, daughter or caregiver. Each record of the Incident File can be thought of as representing a number of victimization incidents experienced by persons in the overall population. This number is given by the weighting factor WGHT_VIC. Usually there is a report for each victimization incident, but victimization incidents with very similar details are recorded on the same report (known as a series report). The number of incidents that the report represents is known as the series factor and is given by variable NUMINC. To estimate the total number of incidents with a given characteristic, one would multiply WGHT_VIC by the series factor and sum over all records with the characteristic. For example, to estimate the total number of break and enter incidents that were reported to the police, WGHT_VIC is multiplied by NUMINC and the resulting value is summed over all records representing a break and enter incident that was reported to the police.

Note that some series reports involve a large number of similar incidents. Some analysts may feel that leaving them as is will imply a disproportionate contribution to victimization estimates from this type of incident. Indeed, the series factor was capped at 3 for estimates published in this user's guide. If analysts wish to use the same cap for the series factor, they may use the weighting factor ADJWTVIC, which is WGHT_VIC multiplied by the capped series factor. In the above example, to estimate the total number of break and enter incidents reported to police, with any series reports capped at 3, ADJWTVIC would be summed over all records representing a break and enter incident that was reported to the police.

Note also that violence by a current or ex-spouse and violence against seniors by their children or caregivers is only captured in abuse reports on the Main File. Suppose a given estimate of a number of victimization incidents is to include violence of this type. Then the number of victimization incidents involving this type of violence must be calculated from the Main File separately and then added to the estimate from the Incident File. More information and examples about these kinds of estimates can be found in Section 7.6.

Weighting of Persons and Scenarios

We view each cycle of the General Social Survey as being composed of a number of independent surveys - one per collection month. Wherever possible, therefore, we weight each monthly survey independently so that the data collected for each month contribute to the estimates in proportion to the Canadian population for that month. Where the sample size for a particular month is not large enough, the records for two or more months are grouped together at certain stages of the weighting process.

A self-weighting sample design is one for which the weights of each unit in the sample are the same. The GSS sample for Cycle 13 was selected using the Elimination of Non-Working Banks (ENWB) sampling technique, which has such a design, with each telephone number within a stratum having an equal probability of selection.

This probability is equal to:

$$\frac{\text{Number of telephone numbers sampled within the stratum}}{\text{Total number of possible telephone numbers within the stratum}}$$

Total number of possible telephone numbers within the stratum

(The total number of possible telephone numbers for a stratum is equal to the number of working banks for a stratum times 100).

1) Basic Weight Calculation

Each working (in service) telephone number (responding and non-responding) in the RDD sample was assigned a weight equal to the inverse of its probability of selection. This weight was calculated independently for each stratum-month group as follows:

$$\frac{\text{Number of possible telephone numbers in each stratum-month group}}{\text{Number of sampled telephone numbers in each stratum-month group}}$$

Number of sampled telephone numbers in each stratum-month group

2) Non-Response Adjustment

Weights for responding telephone numbers were adjusted to represent non-responding telephone numbers. This was done independently within each stratum-month group. The basic weights were adjusted by multiplying by the following factor:

$$\text{Factor 1} = \frac{\text{Total of the basic weights of all telephone numbers in each stratum-month group}}{\text{Total of the basic weights of responding telephone numbers in each stratum-month group}}$$

Non-responding telephone numbers were then dropped.

3) Household Weight Calculation

The weight from Step 2 was used as an initial household weight. For households with more than one residential telephone number (i.e. not used for business, fax or computer purposes only), this weight was adjusted downwards to account for the fact that such households had a higher probability of being selected. The weight for each household was divided by the number of residential telephone numbers that serviced the household.

$$\text{Factor 2} = \frac{1}{\text{Number of non-business telephone numbers}}$$

This produces a household weight = Basic Weight * Factor 1 * Factor 2.

4) Person Weight Calculation

A person weight was then calculated for the respondent by multiplying the household weight by the number of persons in the household who were eligible to be selected for the survey (i.e. the number of persons 15 years of age or older).

This produces a person weight = Basic Weight * Factor 1 * Factor 2 * Number of eligible household members.

5) Adjustment of Person Weight 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 produced that should match certain external reference totals do indeed match them. Two sets of external references were used for this survey, both of them population totals: for stratum by month, and for age-sex groups by province.

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

5a) Regional Office (RO) - Stratum - Month Adjustment

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

$$\frac{\text{Projected population count for the RO-stratum-month}}{\text{Sum of the person weights for the RO-stratum-month}}$$

When sample sizes were small, adjacent months' data for the same stratum were combined before this adjustment was made.

5b) Province - Age - Sex Adjustment

The next weighting step was to ratio adjust the weights to 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:

$$\frac{\text{Projected population count for the province-age-sex group}}{\text{Sum of the person weights of records for the province-age-sex group}}$$

where,

$$\text{Projected population count} = \frac{\text{Dec 1999} \sum \text{Projected population count for province-age-sex group}}{\text{Feb 1999}}$$

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 RO-Stratum-Month and Province-Age-Sex totals would agree with the projections. 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 5) is the final person weight WGHT_PER placed on the Main File.

7) Scenario Weight:

Respondents were asked a series of four questions about their views on sentencing in a hypothetical crime situation. There were four possible such situations, and one of these was chosen at random to present to the respondent, as identified in the QUSERIES variable. In addition, for each situation, there were two possible orders in which to ask the questions. Again, one of the two orders was assigned at random. The respondents sharing a scenario, that is the same combination of crime situation (i.e. QUSERIES) and order of questions, constitute a sub-sample and there were, of course, eight such sub-samples. Weights were produced for each of these eight sub-samples in a way that was completely analogous to the process described above in 1) to 6). Pairs of sub-samples that differed only in the order that the questions were asked were then combined by dividing the weights by 2, so that each order had equal importance. The resulting weight is the final scenario weight WGHT_SCE placed on the Main File.

8) Household Weight:

The person is considered the basic unit of analysis and many steps have been taken to ensure that the person weight is as accurate as possible, as can be seen above in its derivation. Sometimes, however, estimates of the number of households will be required, so an approximation of a household weight has been included on the Main File. It is simply the person weight divided by the number of household members eligible for selection in the survey, i.e. those aged 15 or over. The latter variable is called DVELLIG:

$$\text{WGHT_HSD} = \text{WGHT_PER} / \text{DVELLIG}^3$$

³ DVELLIG is not included on the microdata files.

Note that the adjustments performed on the person weight such as the raking ratio make WGHT_HSD preferable to the preliminary household weight mentioned in 3) above.

9) Abuse Weight

Victimization incidents that occurred within certain types of violent relationships were not examined individually; instead, respondents were asked about the violent relationship as a whole. An abuse report was filled out rather than several incident reports. This was true for spousal and ex-spousal violence, and violence against seniors involving a son, daughter or caregiver. The number of incidents having occurred in these types of violent relationships over the past 12 months was recorded, however. (They could thus be included in any estimates of total number of crime incidents having occurred in this time period.) The weighting factor WGHT_ABU should be used for estimates of numbers of incidents from the above types of violent relationships. It was created to stress that numbers of incidents are being estimated, but it is the same as the person weight:

$$\text{WGHT_ABU} = \text{WGHT_PER}$$

7.2 Weighting of Victimization Incidents

The victimization incidents were weighted after the person weights had been produced. The final victimization weight, WGHT_VIC, was calculated from the person weight, WGHT_PER, following the procedures given below.

Each of 10,087 victimization reports was classified as being either a "person" or "household" crime based on the value of MSCRIME. MSCRIME is a derived variable for the most serious crime included on a given report.

Weighting Person Victimization Incidents

For various reasons, **the definition of a victimization incident for Cycle 13 has changed from that of past victimization cycles.** In these past cycles, a person crime event that involved several victims was counted as one incident. In Cycle 13, however, it was defined as several incidents, one incident per victim. Reasons for the change include:

- 1) It is difficult to talk about the demographic characteristics of the victim of the incident when there is more than one victim.
- 2) The new "one victim, one incident" rule is more consistent with police charges.
- 3) The old definition required knowing how many other people were victims during the incident (see below). This information was provided by the respondent and was rather subjective. Indeed, the same crime might be viewed as several incidents by one respondent but as only one by another.

The change is emphasized by the fact that the weight for incident reports is now known as a

victimization weight – WGHT_VIC (rather than WGHT_CRI).

Analysts are therefore cautioned against making comparisons between Cycle 13 and previous victimization cycles without adjusting the definition of crime incident so that it's consistent among cycles.

For Cycle 13, the victimization weight for person crimes is simply the person weight, so:

$$\text{WGHT_VIC} = \text{WGHT_PER}$$

(In past victimization cycles, one accounted for the probability that the person crime incident could have been reported by the other persons harmed or threatened in the incident. The person weight was divided by the variable representing the number of persons 15 years of age or older victimized in the incident.)

Weighting Household Victimization Incidents

The definition of a household victimization incident is the same as that of household crime incident in previous victimization cycles. Household crimes could have been reported by any eligible member of the sampled household so an adjustment to the person weight was made to account for this. The variable DVELLIG is the number of household members who are 15 years old or older.

The household victimization incident weight, WGHT_VIC, is then calculated as:

$$\text{WGHT_VIC} = \text{WGHT_PER} / \text{DVELLIG}^4$$

IMPORTANT: Each of the Victimization Incident Reports can represent either a single victimization incident or a series of victimization incidents. See Section 7.6 for more information on this.

7.3 Weighting Policy

Users are cautioned against releasing unweighted tables or performing any analysis based on unweighted survey results. As was discussed in Section 7.1, there were several weight adjustments 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, it is known that non-respondents are more likely to be males and more likely to be younger. In the responding sample, 3.2% were males between the ages of 15 and 19, while in the overall population, approximately 4.3% were males between 15 and 19. Therefore, it is clear that unweighted sample counts cannot be considered to be

⁴ DVELLIG is not included on the microdata files.

representative of the survey target population.

Contact was made or attempted with 31,878 households during the survey. From these households, 25,876 usable responses were obtained, for a response rate of 81.2%⁵ (when it is assumed that all of the households for which there was no response were "in scope", i.e., had at least one eligible member). The distribution of types of non-response and response is shown in the table below:

Total sample of households	31,878	100%
1 Households not reached	950	3.0%
2 Household refusal	2,108	6.6%
3 Other hhld non-response	853	2.7%
4 Selected person refusal	905	2.8%
5 Other person level non-response	1,186	3.7%
6 Responses	25,876	81.2% ⁶

Lines 1, 2, and 3 above represent non-response that occurred at the household level; in total there were 3,911 household non-responses, 12.3% of the sample. Line 1 indicates the number of households that could not be reached during the entire survey period ("ring-no-answer" households). Lines 4 and 5 represent non-response that occurred after the respondent for the household had been selected. In total there were 2,091 of these person level non-responses, 6.6% of the sample. The 'other non-response' categories include cases where a response could not be obtained due to language difficulties or other problems.

7.4 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 or averages. More complex estimation and analyses are covered in Section 7.5.

⁵ This is the response rate AFTER processing. The response rate observed in the field during collection was 81.3%. The discrepancy is due to the fact that a few cases changed status during processing.

⁶ This is the response rate AFTER processing. The response rate observed in the field during collection was 81.3%. The discrepancy is due to the fact that a few cases changed status during processing.

7.4.1 Qualitative Estimates

It should be kept in mind that the target population for the GSS was non-institutionalized persons 15 years of age or over, 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 (6,301,805) who describe their state of health as excellent (HLTHSTAT=1, derived from Q31A) is an example of this kind of estimate. These estimates are readily obtained by summing the person weights (WGHT_PER) of 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 ignoring the respondents with a 'Not stated' or 'Don't know' answer to question Q31A and calculating the ratio of the total of the weights of those respondents who answered that their state of health was 'excellent' (Q31A=1). 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):

$$6,679,657 = 24,260,326 \times \frac{6,301,805}{22,887,978}$$

When the proportion of responses that are 'don't know' or 'refused' is high the differences between the two estimates will be large.

Another example of a qualitative estimate is the number of people (4,165,026) who regularly have trouble going to sleep or staying asleep (SLEEPROB = '1', derived from Q30). Again this estimate does not adjust for non-response to the question in any way. The adjustment is done and a final estimate produced by following the same method used in the previous example. We end up with the final estimate being:

$$4,379,747 = 24,260,326 \times \frac{4,137,108}{22,916,300}$$

7.4.2 Quantitative Estimates

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

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

The numerator (X) is a quantitative estimate of the total of the variable of interest (for example, the number of weeks worked in the past 12 months) for a given sub-population (for example, males in Ontario 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 (Q36) when it is known, $1 \leq Q36 \leq 52$, (i.e. not equal to '96', '97' or '99'), and summing this product over all records for males in Ontario who worked i.e. SEX=1 and PRV=35 and ($1 \leq Q36 \leq 52$), which yields 141,739,791.

The denominator (Y) is the qualitative estimate of the number of persons within that sub-population (males in Ontario 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 in Ontario with $1 \leq Q36 \leq 52$, yielding 2,792,238.

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:

$$\frac{141,739,791}{2,792,238} = 50.8$$

7.5 Guidelines for Analysis

As is detailed in Section 4 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 and multiple stages of selection, and unequal probabilities of selection of respondents.

Using data from such complex surveys presents problems to analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used.

The GSS used a stratified design, with significant differences in sampling fractions between strata. Thus, some areas are over-represented in the sample (relative to their populations) while some other areas are relatively under-represented; this means that the unweighted sample is not representative of the target population, even if there were 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, with the result that while in many cases the estimates produced by the packages are correct, the variances that are calculated are almost meaningless.

For many analysis techniques (for example linear regression, logistic regression, 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 on the data, or on 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.

For an analysis of all respondents who were victimized in the past 12 months (including spousal violence and violence against seniors), the following steps are required:

- Select all respondents from the file who were victimized in the past 12 months (TOTVIC_X > 0);
- Calculate the Average Weight of WGHT_PER for these records;
- For each of these respondents calculate a "working" weight equal to WGHT_PER / Average Weight;
- Perform the analysis for these respondents using the "working" weight.

Section 8 describes sampling variability and data reliability in more detail and Appendix A gives a series of tables that can be used to estimate the sampling variability of many qualitative estimates of totals and proportions.

7.6 Methods of Estimation and Interpretation of Estimates

7.6.1 Estimating Numbers 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 projected by Statistics Canada for each month covered by Cycle 13.

$$\begin{aligned} 25,876 \\ \sum_{i=1} \text{WGHT_PER} &= 24,260,326 \\ &= \text{an estimate of the number of persons 15 years of age and} \\ &\quad \text{older in the population.} \end{aligned}$$

In general, when an estimate is based on the unit of observation being the person, the Main File and WGHT_PER should be used. Examples of this are the average number of weeks worked by persons aged 25-29, the percentage of persons aged 65 or over who feel that crime in their neighbourhood has increased, and the number of people aged 15-19 who have been victim of at least one crime (of a type included in the survey) in the 12 months prior to the survey.

The last example would be calculated as follows. WGHT_PER would be summed up for all records on the main file with $1 \leq \text{AGEGR5} \leq 2$ and $\text{TOTVIC_X} > 0$ to find 822,076 people aged 15-19 who have been victim of at least one crime in the 12 months prior to the survey.

7.6.2 Estimating Numbers of Victimization Incidents by Using WGHT_VIC or ADJWTVIC on the Incident File:

In addition, as previously mentioned, a victimization weight has been assigned to each incident report. Again as previously mentioned, incidents that occurred in certain violent relationships are not given individual reports on the Incident File. Such relationships involve violence by a current or ex-partner or violence against a senior by a child or caregiver. Finally, some reports may represent more than one incident, in which case the report is known as a series report. The number of incidents represented by the report is known as the series factor and is available in variable NUMINC.

10,087

$$\sum_{i=1} \text{NUMINC} * \text{WGHT_VIC} = 10,753,527.$$

= an estimate for the number of victimization incidents excluding violence by a current or ex-partner and violence against a senior by a child or caregiver.

It may well seem reasonable to not give too much importance to series reports, some of which involve a very large number of incidents. One possible approach is to count them as only one incident, in which case we get:

10,087

$$\sum_{i=1} \text{WGHT_VIC} = 6,751,896.$$

= an estimate for the number of victimization incidents excluding violence by a current or ex-partner and violence against a senior by a child or caregiver.

A compromise between the above two approaches is to cap the series factor at 3, i.e. define a new variable, say NUMINCCA, in the following way:

```
IF NUMINC > 3 THEN
  NUMINCCA = 3
ELSE
  NUMINCCA = NUMINC.
```

10,087

$$\sum_{i=1} \text{NUMINCCA*WGHT_VIC} = 7,608,212.$$

= an estimate for the number of victimization incidents excluding violence by a current or ex-partner and violence against a senior by a child or caregiver.

This is the approach that was used for all published estimates of victimization incidents in this user's guide. For the user's convenience when he or she wishes to follow this approach, the pseudo-weight ADJWTVIC has been placed on the incident file. $\text{ADJWTVIC} = \text{NUMINCCA*WGHT_VIC}$. We therefore have:

10,087

$$\sum_{i=1} \text{ADJWTVIC} = 7,608,212.$$

= an estimate for the number of victimization incidents excluding violence by a current or ex-partner and violence against a senior by a child or caregiver.

In general, when an estimate is based on the unit of observation being a victimization incident, the Incident File and WGHT_VIC should be used. Examples of this are the number of break and enter incidents reported to the police by the respondent, the number of incidents of vandalism with the cost of the damage exceeding \$1000, and the percentage of physical assault incidents involving more than one assailant. However, any characteristics of the victim in a person crime will have to be retrieved from the Main File by linking with the record identifier (RECID). For example, this will be the case if the number of personal property theft incidents involving young victims from Ontario aged 15 to 24 is desired. Furthermore, as has been mentioned, the Main File must be used for estimation of numbers of incidents of spousal violence and violence against seniors by their children or caregivers.

Here is how two of the above examples are calculated. To estimate the number of break and enter incidents reported to the police by the respondent, capping any series reports at 3, ADJWTVIC would be summed up for all records on the Incident File with MSCRIME = 405 and V53 = 1, yielding 217,404 incidents. To estimate the number of incidents of property theft involving young victims from Ontario aged 15 to 24, RECID on the Incident File would be compared to RECID on the Main File and variables AGEGR5 and PRV from the latter retrieved when there's a match. ADJWTVIC would be summed up for all records on the now "augmented" incident file with MSCRIME = 609, PRV = 35 and $01 \leq \text{AGEGR5} \leq 03$, yielding 172,938 incidents.

7.6.3 When Estimating Numbers of Persons Will Require the Use of the Incident File

The above generalizations regarding which file and weight to use do not apply to all estimates. Some characteristics of the respondent that are needed for person based estimates can only be determined from the Incident File. Since estimates derived from such characteristics are based on the person as the unit of observation (rather than the incident), the weight WGHT_PER should be used rather than WGHT_VIC, even though the Incident File is used.

For instance, the number of persons 15 years of age or older who have been victim of a crime within the past 12 months of the survey can be estimated from the Main File (sum WGHT_PER for all records with TOTVIC_X greater than 0, yielding 6,291,803). However, the number of persons 15 years of age and older who were victim of one or more crimes related to the offender's alcohol or drug use can only be estimated using the Incident File. This is because, while the estimate is for a person level characteristic, the person level file, the Main File, does not have the necessary detail about each victimization incident. In essence, you have to use the Incident File to derive a person level variable (as opposed to an incident level variable), that you then use with WGHT_PER to produce the estimate of interest. In this case, you would derive a variable that indicated the existence, among all of the incidents (if any) on the Incident File for a respondent, of one or more that were related to the offender's alcohol or drug use, and then sum WGHT_PER over all of those persons where the variable indicated at least one such incident. (For the sake of simplicity, we will assume that the incidents of interest exclude spousal violence and violence against seniors by their children or caregivers.)

To calculate the estimate described above, one would create a new variable that indicated the existence of an incident related to the offender's alcohol or drug use. This variable would be set to the value meaning 'No' when one encounters the first incident for a respondent, then each incident for that respondent would be examined in turn from the first to the last. If an incident was related to the offender's alcohol or drug use (V20A = 1), then the new variable would be set to 'Yes'. If after examining the last episode for a respondent the new variable has a value of 'Yes' then WGHT_PER for that respondent would be added to the estimate. In this example, the desired estimate works out to 899,258 persons.

7.6.4 When Estimating Numbers of Victimization Incidents Will Require the Use of the Main File

As has been repeatedly stressed, incidents on the Incident File exclude spousal violence and violence against seniors by their children or caregivers. For estimates to include incidents of this type, two separate calculations must be done and then added, i.e. the contribution from the Main File (incidents from spousal violence and violence against seniors by their children or caregivers) and the contribution from the Incident File (other incidents). On the Main File, the numbers of incidents that a given respondent has experienced are found in variables PR_101, PR_304, SN_101 and SN_304 (see section 6.4.2) and should be weighted by WGHT_ABU.

As an example, suppose an estimate is desired of the total number of incidents of sexual and physical assault (with series capped at 3). ADJWTVIC on the Incident File would be summed up over all records with MSCRIME in (101,304), yielding 1,745,527. Adding PR_101, PR_304, SN_101 and SN_304 gives, for any particular respondent, the number of incidents of spousal violence and violence against seniors by their children or caregivers. Multiplying it by WGHT_ABU and summing over all Main File records gives 717,800, an estimate of the total number of incidents of this type of violence. Adding the two figures together yields 2,463,327, an estimate of the total number of incidents of sexual and physical assault (with series capped at 3).

7.6.5 Estimating Numbers of Persons by Using WGHT_SCE on the Main File

The scenario weights should be used when estimates involve people's opinions on the sentencing of offenders. Recall that the variable QUSERIES identifies which of the four possible hypothetical situations was presented to a given respondent (see Section 6.4.1).

The scenario weights should be used analogously to the person weight for estimates involving variables SCEN_A, SCEN_AP, SCEN_Y and SCEN_YP. However, **please note that these variables contain the answers to different questions depending on which situation was presented to the respondent. They will give meaningless results unless QUSERIES is specified.** For example, consider the number of persons 15 years or older who have been victim of a crime in the past 12 months and who feel that a prison sentence is appropriate for a first-time adult offender found guilty of breaking into a house while the owners are on vacation and taking goods worth \$400. WGHT_SCE would be summed over those with QUSERIES=A and SCEN_A=1 and TOTVIC_X > 0, yielding an estimate of 2,459,306.

Note: It is possible to use WGHT_SCE in order to obtain estimates for variables not related to the scenarios. For example, summing WGHT_SCE over all records with QUSERIES=A and TOTVIC_X > 0 yields 6,187,342, an estimate of the number of persons who were victim of at least one crime in the past 12 months. Doing the same for records with QUSERIES=B, C and D yields other estimates, namely 6,442,938, 6,113,829 and 6,370,433, respectively. Summing WGHT_PER over all records with TOTVIC_X > 0 yields 6,291,803. The latter is the estimate to use, as it is obviously better to take the whole sample rather than a quarter of the sample. As has already been stressed, WGHT_SCE should only be used for scenario questions.

7.6.6 Estimating Numbers of Households by Using WGHT_HSD on the Main File

The weighting factor WGHT_HSD should be used when estimating the number of households. For example, when estimating the number of households in Newfoundland, WGHT_HSD would be summed over all records with PRV=10, yielding an estimate of 202,409 households.

8. 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 microdata files.

This section of the documentation provides guidelines to be followed by users. With the aid of these guidelines, users of the microdata files 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 can be broken into four broad sections: Minimum Sample Sizes for Estimates; Sampling Variability Policy; Sampling Variability Estimation; and Rounding Policy.

8.1 Minimum Sample Size For Estimates

Users should determine the number of records on the particular microdata file which contribute to the calculation of a given estimate. This number should be at least 15 in the case of persons and households and at least 40 in the case of victimization incidents. When the number of contributors to the weighted estimate is less than this, the weighted estimate should generally not be released regardless of the value of the Approximate 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.

8.2 Sampling Variability Guidelines

The estimates derived from this survey are based on a sample of households. 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 implemented at each step of the data collection and processing cycle to monitor the quality of the data. These measures included the use of highly skilled interviewers, extensive training of interviewers with respect to the survey procedures and questionnaire, observation of interviewers to detect problems of questionnaire design or misunderstanding of instructions, procedures to ensure that data capture errors were minimized and coding and edit quality checks to verify the processing logic.

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 just 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.

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 with certainty that the differences would be less than three standard errors.

Because of the large variety of estimates that can be produced from a survey, the standard error is usually expressed relative to the estimate to which it pertains. The resulting measure, known as the coefficient of variation (c.v.) of an estimate is obtained by dividing the standard error of the estimate by the estimate itself and is expressed as a percentage of the estimate. Before releasing and/or publishing any estimates from the microdata file, users should consider whether or not to release the estimate based on the following guidelines.

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.

Note: The sampling variability guidelines should be applied to rounded estimates.

8.3 Estimates of Variance

Variance estimation is described separately for qualitative and quantitative estimates.

8.3.1 Sampling Variability for Qualitative Estimates

Derivation of the sampling variability of each of the qualitative estimates which could be generated from the survey would be an extremely costly procedure, and for most users, an unnecessary one. Consequently, approximate measures of sampling variability, in the form of tables, have been developed for use and are included in APPENDIX A ("Approximate Variance Tables"). These tables were produced using the coefficient of variation formula based on a simple random sample. Since the design of Cycle 13 of the General Social Survey was not a simple random sample, a factor called the Design Effect has been introduced into the variance formula.

The Design Effect for an estimate is the actual variance for the estimate (taking into account the design that was used) divided by the variance that would have resulted if the estimate had been derived from a simple random sample. The Design Effect used to produce the Approximate Variance Tables has been determined by first calculating Design Effects for a wide range of characteristics and then choosing among these a conservative value which will not give a false impression of high precision. These Design Effects are specified in the table that follows.

Design Effects

<u>Geographic Area</u>	<u>Person Design Effect</u>	<u>Scenario Design Effect</u>	<u>Victimization Design Effect</u>
Canada	1.41	1.39	4.75
Newfoundland	1.52	1.70	4.48
Prince Edward Island	1.12	1.29	3.33
Nova Scotia	1.20	1.24	2.77
New Brunswick	1.24	1.16	2.68
Quebec	1.27	1.23	3.02
Ontario	1.33	1.30	4.28
Manitoba	1.32	1.34	4.24
Saskatchewan	1.20	1.17	3.52
Alberta	1.14	1.16	3.07
British Columbia	1.19	1.28	4.22
Atlantic Region	1.52	1.47	3.34
Prairie Region	1.23	1.15	3.58

Approximate variance tables for estimates using WGHT_PER and ADJWTVIC (i.e. with series reports capped at 3) are provided at the Canada and provincial levels as well as for the Atlantic and Prairie Regions. Approximate variance tables for estimates using WGHT_SCE are provided at the Canada level.

It should be noted that all coefficients of variation in these tables are approximate and therefore unofficial. Variable specific estimates of variance calculated for particular variables may be purchased from Statistics Canada. The use of variable specific variance calculation instead of the table-based approximations may allow users to feel more certain of the quality of their estimates, especially those with coefficients of variation estimated from the tables in the "Very High Sampling Variability" range (see the guidelines regarding the release of the survey estimates on preceding pages).

Statistics Canada is investigating the feasibility of releasing to GSS microdata file users a set of supplementary weights that would allow them to calculate a variable specific variance for any estimate produced from the microdata file. The variance calculation would be done using the bootstrap method. A large number of additional weights, known as bootstrap weights, would be provided for each respondent. When a variable specific variance estimate is required, the estimate for the variable in question would be first made with the standard weight, and then using each of the bootstrap weights in its place to produce many bootstrap versions of the same estimate. The variance of the set of bootstrap estimates can be used to calculate an estimate of the sampling variability of the estimate of interest. Please contact Statistics Canada for more information on the availability of the bootstrap weights and on the bootstrap method for the calculation of variable specific variance estimates by microdata file users.

8.3.2 Sampling Variability For Quantitative Estimates

Approximate variances for quantitative variables cannot be as conveniently summarized. As a general rule, however, the coefficient of variation of a quantitative total will be larger than the coefficient of variation of the corresponding qualitative estimate (e.g., the number of persons contributing to the quantitative estimate). If the corresponding qualitative estimate has high sampling variability, then the quantitative total will in general have high sampling variability as well.

8.4 Rounding

In order that estimates produced from the General Social Survey microdata files 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.

8.4.1 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 8.4.2).

- 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.

8.4.2 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.

9. ADDITIONAL INFORMATION

Additional information about this survey can be obtained from the individuals listed below. Data from the survey is available through published reports, special request tabulations, and this micro data file. The microdata file is available from the Housing, Family and Social Statistics Division of Statistics Canada at a cost of \$1,600. Tabulations can be obtained at a cost that will reflect the resources required to produce the tabulation.

Subject Matter

Karen Mihorean
Canadian Centre for Justice Statistics
(613) 951-4496
Karen.Mihorean@statcan.ca

Sample Selection Procedures, Weighting and Estimation

Paul Matthews
Household Survey Methods Division
(613) 951-1480
Paul.Matthews@statcan.ca

Data Collection and Data Processing

Nancy Zukewich
Housing, Family and Social Statistics Division
(613) 951-8645
Nancy.Zukewich@statcan.ca