

Table of Contents

1.	Introduction	1
2.	Objectives of GSS	1
3.	Target Population	2
4.	Survey Design.....	3
5.	Collection.....	4
6.	Processing	6
7.	Estimation.....	8
8.	Release Guidelines and Data Reliability.....	23
9.	File Structure	29
10.	Additional Information	31
Appendix A.	Approximate Variance Tables	A1 - A21
Appendix B.	Survey Questionnaire	B1 - B93
Appendix C.	Topical Index to Variables for Main File	C1 - C23
Appendix D.	Data Dictionary for Main File	D1 - D404
Appendix E.	Record Layout for Main File	E1 - E18
Appendix F.	Data Dictionary for Time Use Episode File	F1 - F16
Appendix G.	Record Layout for Time Use Episode File	G1 - G3
Appendix H.	1998 Activity Coding List and Instructions	H1 - H62
Appendix I.	1992 Activity Coding List	I1 - I4
Appendix J.	1998 to 1992 Activity Code Comparison	J1 - J10
Appendix K.	1992 to 1998 Activity Code Comparison	K1 - K10
Appendix L.	1998 Twenty-four Code Activity System	L1 - L8
Appendix M.	1998 GSS Sports Code List	M1 - M2
Appendix N.	A guide to using Time Use Data	N1 - N13

1. INTRODUCTION

This document is designed to enable interested users to access and manipulate the microdata file for the twelfth cycle of the General Social Survey, conducted from February 1998 through January, 1999. It contains information on the objectives, methodology and estimation procedures as well as guidelines for releasing estimates based on the survey.

This document gives a description on how to correctly use the microdata files. Appendices D and F contain the data dictionaries for the Main File and the Time Use Episode File, respectively, which is the major part of this documentation package. The variance tables are in Appendix A and the survey questionnaire, in Appendix B.

This package is available in machine readable form.

2. OBJECTIVES OF THE GENERAL SOCIAL SURVEY

Increased pressure, over the last decade, to operate more efficient government funded programmes has led to a related increase in the information needed for policy formulation, programme development and evaluation. Many of these needs could not be filled through existing data sources or vehicles because of the range or periodicity of the information required.

The two primary objectives of the General Social Survey (GSS) aim at closing these gaps. These objectives are: to gather data on social trends in order to monitor temporal changes in the living conditions and well-being of Canadians; and to provide immediate information on specific social policy issues of current or emerging interest. The GSS is a continuing program with a single survey cycle each year. However, there was no GSS undertaken for the 1997 reference year due to budgetary priorities at that time.

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 is designed to obtain information which monitors social trends or measures changes in society related to living conditions or well-being. Cycle 12 was the third cycle to return to previous core content: time use. Most of the core content of Cycle 12 repeated Cycles 7 and 2, conducted in 1992 and 1986, respectively.

Focus content is aimed at the second survey objective of GSS. This component obtains information on specific policy issues which are of particular interest to certain federal departments or other user groups. In general, focus content, is not expected to be repeated on a periodic basis. The focus content for Cycle 12 was participation in sport and cultural activities sponsored by Heritage Canada. Information on media use was also collected for the Canadian Broadcasting Corporation.

3. TARGET POPULATION

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

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

In the survey, all respondents were contacted by telephone. Households without telephones were therefore excluded; however, persons living in such households represent less than 2% of the target population. Survey estimates have been adjusted (i.e., weighted) to account for persons 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. SURVEY AND SAMPLE DESIGN

Data for Cycle 12 of the GSS were collected monthly from February 1998 to January 1999 inclusive. The sample was evenly distributed over the 12 months to represent the seasonal variation in the information. The sample was selected using the Elimination of Non-Working Banks technique of Random Digit Dialing (RDD). Since people's activities also differ by the day of the week, a sample that was representative of each day of the week was required. Each telephone number was therefore assigned a designated day. Cases were eligible for collection for 2 days following the designated day; with priority given to collecting diary information on the day following the designated day.

The response rate for Cycle 12 was 77.6%, yielding 10,749 respondents from whom usable diary information was obtained. A description of the RDD methods is provided in Section 4.2. Stratification used in the survey design is outlined in Section 4.1.

4.1 Stratification

In order to carry out sampling, each of the ten provinces was divided into strata or geographic areas. Generally, for each province, one stratum represented the Census Metropolitan Areas (CMAs) of the province and another represented the non-CMA areas.

There were two exceptions to this general rule:

- Prince Edward Island has no CMAs and so did not have a CMA stratum
- Montreal and Toronto were each separate strata.

There were small changes from previous GSS cycles in the allocation of the sample to the various strata because it is based on the total populations of the strata, and these change slightly, relative to each other, every year.

4.2 Elimination of Non-working Banks RDD Design

The Elimination of Non-Working Banks (ENWB) sampling technique is a method of Random Digit Dialing 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 whom 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 had to be made as to the percentage of numbers dialed that would reach a household (this percentage is known as the "hit rate"). Hit rates from Cycle 11 RDD sample and those observed during data collection were used to estimate the hit rates for the Cycle 12 RDD sample monthly samples.

For Cycle 12 of the GSS, 46.8% of the numbers dialed reached households. An attempt was made to conduct a GSS interview with one randomly selected person from each household.

5. COLLECTION

As in the other General Social Surveys taken since 1994, data for Cycle 12 were collected using Computer

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.

Assisted Telephone Interviewing (CATI) using Computer-Assisted Survey Execution System software (CASES). With CATI, the survey questions appeared on a computer monitor. The interviewer asked the respondent the questions, and entered the responses into the computer as the interview progressed. CATI methodology eliminated the need for paper and pencil questionnaires.

All interviewing took place using centralized telephone facilities in four of Statistics Canada's regional offices with calls being made from 9:00 until 21:00, Monday to Friday inclusive, and from 12:00 until 16:00 on Saturday and Sunday. 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 in a two day classroom training session. The majority of interviewers had computer and telephone interviewing experience.

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 automatically followed the flow of the questionnaire. Certain edits were also executed by the CATI system. The data were then transmitted to Ottawa electronically.

In Cycle 12, the CATI system provided the interviewer with two main "components" which can be imagined to represent two paper questionnaires:

GSS 12-1 Selection Control Questionnaire

GSS 12-2 Time Use Questionnaire

These are included in the document *A Cycle 12 Questionnaire Package*.[@] The questionnaire documents include the skip patterns that were built into CATI although these did not appear on the screen.

A GSS 12-1 selection control questionnaire was completed for each telephone number generated in the sample. When a private household was contacted, all household members were enumerated and basic demographic information (age, sex, marital status) was collected for everyone. A computer algorithm randomly selected an eligible household member age 15 or over to answer the Time Use questionnaire. This form was also used to determine the eligible collection days for the purpose of scheduling appointments to complete the questionnaire.

The Time Use questionnaire collected the following types of information: general questions related to time

(Section A); the time use diary (Section B); a child care diary for respondent's with children less than 15 years of age living in the household (Section C); perceptions of time (Section D); information on unpaid help supplied by the respondent to household members and other individuals, as well as information on volunteering (Section E); paid work and education activities of the respondent (Section F); work and education activities of the respondent's partner or spouse, if applicable (Section G); cultural and recreational activities of the respondent (Section H); sport participation (Section J); enjoyment of activities (Section K); background socio-economic questions for classification purposes (Section L).

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 10). Shown below is a list of the manuals used in the survey:

GSS Cycle 12 Time Use Interviewer's Manual

GSS Cycle 12 Time Use Coding 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 the Head Office in Ottawa.

6.2 Coding

The coding of the daily activities was done by the interviewer during the interview with the aid of CATI screens. These formed a decision tree leading to a narrow selection of possible codes. These screens are included in the questionnaire package.

Diary activities that the interviewer could not code (specified) were coded manually at head office from write-in information. In addition, write-in information from residual activities (specified) was provided on

the diary to describe selected types of activities. These were reviewed at head office and where necessary, the code was changed. In some instances, large numbers of similar activities were observed in a residual activity code and new codes were created.

This activity coding methodology was very different from that used in previous time use surveys. In Cycle 7, for example, the senior interviewer coded diary entries on a paper questionnaire, using a coding manual. During the collection of Cycle 12, selected diaries were concurrently captured by paper and pencil. A comparison of the coded results will be done to evaluate the impact of the new CATI coding method.

Where possible (e.g., occupation, industry, education, country of birth and religion), the coding followed the standard classification systems as used in the Census of Population. In addition, the write-in information describing responses in residual codes was reviewed: when appropriate, some were coded to existing categories. In some instances new codes were created. All interviewers comments were reviewed and taken into account in head office editing.

6.3 Edit and Imputation

During the interview, the CATI system ensured that branchings were correct and that the values were valid. In cases where the interviewer was unable to correct errors detected by the system, he/she could provide a comment and leave the problem for Head Office to solve.

The Head Office edit system performed the same kind of checks as the CATI system, as well as verifications of greater complexity. The flow of the responses through the various paths in the questionnaire were verified.

The data items used for weighting, such as age, sex and number of telephone lines, needed to have values for all respondents. In the case of the age variable, the procedure used to select the respondent ensured that a response would be present. By contrast, values were imputed in the rare cases where the number of residential telephone lines was missing. The number of residential telephone lines was assumed to be one (1) when the respondent failed to provide the information.

Due to the nature of the survey, imputation was not appropriate for most items and thus 'not stated' codes were usually assigned for missing data. In some cases, the answer was not known but could be obtained

deterministically from other information on the survey.

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.

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 have been aggregated into broader classes (e.g., occupation, religion, country of birth).

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.

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 12 estimates can be made. The **Main File** contains summary time use information from 10,749 respondents. It also contains the questionnaire responses obtained from these respondents. The **Time Use Episode File** contains information describing the details of the 221,105 time use episodes reported by these respondents. Questionnaire information was not collected for those respondents who refused to complete a full diary. For a description of the file layouts, contents and correct interpretation of data on the microdata file, users should refer to Appendices D, E, F, G, H and N.

When analyzing GSS Cycle 12 data, it is necessary to use one of the weighting factors WGHTFIN on the Main File and WGHTEPI on the Time Use Episode File. WGHTFIN indicates the number of persons in the population that a record on the Main file represents, while WGHTEPI indicates the number of time use episodes that a record on the Episode file represents. For example, using the Main File, the estimate of the number of Canadians 15 years of age and older who feel 'trapped in a daily routine' (i.e. D2G = 1) is 8,859,095. This is the sum of WGHTFIN over all records on the Main File with D2G = 1. Using the Time Use Episode File, the estimate of the number of episodes of watching TV by Canadians 15 years of age and older in an average day is 33,559,271, the sum of WGHTEPI over all records on the Time Use Episode File with ACTCODE=911, 912, 913, or 914.

In the last GSS Time Use survey, Cycle 7, there were also two weights, one to be used for estimates not based on the episode data and one for estimates that used the episode data. In Cycle 7 the time use diaries were only collected for a subsample (due to non-response to the diary section of the questionnaire). To account for this subsampling of episodes, the episode weights were larger than the person weights (in the same way that the person weights differ from the household weights to account for the sampling of only one respondent per household (see Section 7.1- 4)). In this GSS cycle, there was no sub-sampling of time use episodes; GSS Cycle 12 collected data for all time use episodes during the reference day for all respondents. Since there was no sub-sampling, the weight for each episode is the same as the weight for the respondent by whom it was reported³.

The Time Use Episode File is structured differently from the Main File in that there are multiple time use episode records for each respondent. Each time use episode is a separately identified record, with each respondent having on average 21 episode records. This introduces additional complexity in applying the weights correctly. Users should refer to Appendix N for the correct methods of using this file.

7.1 Weighting

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

³In the last GSS Time Use survey, Cycle 7, diaries were only collected for a sub-sample (due to non-response to the diary section of the questionnaire) and thus the weight for the Cycle 7 Time Use Episode File differs from that for the Cycle 7 Main File.

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 12 was selected using the Elimination of Non-Working Banks (ENWB) sampling technique, which has such a design, with each household within a stratum having an equal probability of selection.

This probability is equal to:

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

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

Where possible, each survey month was weighted independently. This was done in an attempt to ensure that each survey month contributes appropriately to estimates. If monthly sample sizes were not large enough, two or more survey months were combined in certain steps of the weighting.

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:

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

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. Records were adjusted 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 number⁴ (i.e. not used for business 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 each person who responded to the survey, by multiplying the

⁴Less than 6% of the households in the sample have more than one non-business telephone number.

household weight for that person 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. Three sets of external references were used for this survey, all of them population totals: for stratum by month, for age-sex groups by province, and for day of the week by province by month.

It should be noted that persons living in households without telephone service are included in the external references 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 group-sex distributions. Projected population counts were obtained for males and females within the following twelve 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,

$$\text{Projected population count} = \frac{\text{Jan 1999} \text{ Projected population count for province-age-sex group}}{\text{Feb 1998} \text{ Projected population count for province-age-sex group}}$$

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

5c) Province - Day of the week (Designated Day) - Month Adjustment

Time use information was collected from respondents for a selected day of the week so that each day would have an approximately equal number of respondents. An adjustment was made to the person weights on records within each province, selected day of the week, and month of collection to ensure that population estimates would represent each day of the week. The adjustment was done by multiplying the person weight for each record within the province - day of the week - month combination by the following ratio:

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

where,

$$\text{Projected population count} = \frac{\text{Projected population count for province-month}}{7}$$

5d) 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, Province-Age Group-Sex and Province-Day of the week- Month totals would agree with the projections. This adjustment was made by repeating steps 5a), 5b) and 5c) 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 WGHTFIN placed on the Main File and on the Episode File.

7) Episode Weight

In GSS-7, diaries were collected from only 90% of the respondents who provided otherwise usefully complete data. Thus many time use variables and all the episode data were only available for a

subsample of respondents, so a second weight (TIMEWGT) was needed to account for this subsampling, and this was the weight to use when using the Time Use Episode File. In GSS-12 there was no such subsampling, so the weight that should be used with episode data has the same value as the person weight; it does however have a different interpretation, weighting up to a total over time use episodes rather than over persons, so it has been given a different name, WGHTEPI. It is only on the Time Use Episode File.

7.2 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, sex, and reference day 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, 3.3% 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 representative of the survey target population.

Contact was made or attempted with 13,860 households during the survey. From these households, 10,749 usable responses were obtained, for a response rate of 77.6% (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	13,860	100%
1 Households not reached	613	4.4%
2 Household refusal	1,203	8.7%
3 Other household non-response	337	2.4%
4 Respondent refusal	514	3.7%
5 Other respondent non-response	444	3.2%

6 Responses

10,749

77.6%

Lines 1, 2, and 3 above represent non-response that occurred at the household level; in total there were 2,153 household non-responses, 15.5% 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 958 of these person level non-responses, 6.9% of the sample. The other non-response categories include cases where a response could not be obtained due to language difficulties or other problems.

7.3 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 characteristics) and quantitative estimates involving quantities or averages. More complex estimation and analyses are covered in Section 7.4.

7.3.1 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 (5,461,588) who describe their state of health as excellent (Question L22 = 1) is an example of this kind of estimate. These estimates are readily obtained by summing the final weights (WGHTFIN) 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 'Not stated' or 'Don't know' answer to question L22 and calculating the ratio of the total of the weights of those respondents who answered that their state of health was 'excellent' (L22=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) :

$$5,995,025 = 24,260,137 \times \frac{5,461,588}{\text{-----}}$$

22,101,471

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 women (4,165,026) who are very satisfied with their self-esteem (D6E = >1'). 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:

$$4,494,068 = 12,322,774 \quad \times \quad \begin{array}{r} 4,129,719 \\ \hline 11,323,726 \end{array}$$

7.3.2 Quantitative Estimates

Some variables on the General Social Survey microdata file are quantitative in nature (e.g. age, minutes spent working, 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:

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

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 final weight (WGHTFIN) by the variable of interest (F8 or F13a) when they are known, 1 # F8 # 52 or 1 # F13a # 52, (i.e. not equal to >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 # F8 # 52 or 1 # F13a # 52), which yields 152,570,576.70.

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 final weight (WGHTFIN) over all male respondents in Ontario with 1 # F8 # 52 or 1 # F13a # 52, yielding 3,286,294.16.

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:

$$\frac{152,570,576.70}{3,286,294.16} = 46.4$$

7.4 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, months of the year, or days of the week 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 consider themselves as "workaholics", the following steps are required:

- Select all respondents from the file who considered themselves as a workaholic (D2B = 1);
- Calculate the Average Weight for these records;
- For each of these respondents calculate a "working" weight equal to WGHTFIN / 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.

The calculation of variance estimates that are specific to a variable of interest requires detailed knowledge of the design of the survey; such detail cannot be given in this microdata file because of the risk of breaching confidentiality. Variances that take the sample design into account can be calculated for many statistics by Statistics Canada on a cost recovery basis.

7.5 Methods of Estimation and Interpretation of Estimates

A weight has been assigned to each sampled individual and each sampled time use episode 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 12.

10,749⁵

$$\sum_{i=1}^n \text{WGHTFIN} = 24,260,137$$

i=1

= an estimate of the number of persons 15 years of age and older in the population.

221,105⁶

$$\sum_{i=1}^n \text{WGHTEPI} = 497,943,804$$

i=1

= an estimate for an average day⁷ of the number of time use episodes of all persons 15 years of age and older in the population.

In general, when an estimate is based on the unit of observation being the person, the Main File and WGHTFIN should be used. Examples of this are the average number of weeks worked by persons aged 25-29, the average number of hours spent on housework by persons with children under 5, and the number of persons who watch television on an average day.

Similarly, when an estimate is based on the unit of observation being the time use episode, the Time Use Episode File and WGHTEPI should be used. Examples of this are the average length of an episode of travel for work in the morning, the average length of an episode of meal preparation for the evening meal, and the distribution by time of day of episodes of surfing the net.

However, these generalizations do not apply to all estimates. Some characteristics of the respondent that are needed for person based estimates can only be determined from the Time Use Episode File. Since

⁵There were 10,749 responding households (with one randomly chosen respondent per household).

⁶There were 221,105 time use episodes reported by the 10,749 respondents.

⁷The average over the seven days of the week.

estimates derived from such characteristics are based on the person as the unit of observation (rather than the time use episode), the weight WGHTFIN should be used rather than WGHTEPI, even though the Time Use Episode File is used.

For instance, while the number of Canadians 15 years of age and older who participate in an educational activity on an typical day can be estimated from the Main File (sum WGHTFIN for all records with SCHLEDUC greater than 0, yielding 2,172,602), to estimate the number of Canadians 15 years of age and older who attended educational classes in the evening (after 6:00 PM) on an typical day can only be estimated using the Time Use Episode 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 episode in the respondent's day. In essence, you have to use the Time Use Episode File to derive a person level as opposed to episode level variable that you then use with WGHTFIN to produce the estimate of interest. In this case, you would derive a variable that indicated the existence, among all of the episodes for a respondent, of an educational class that took place after 6:00 PM, and then sum WGHTFIN over all of those persons where the variable indicated the existence of such an activity. Since WGHTFIN is not the correct weight to use with time use episode level data it would not normally be placed on the Time Use Episode File. To facilitate the derivation and use of person level variables from this file, WGHTFIN has been added to the file, but to help avoid its misuse, it has been given a value of blank or missing for all episodes except the last for each respondent.

To make the estimate described above, one would create a new variable that indicated the existence of an educational activity after 6:00 PM. This variable would be set to the value meaning >No= when one encounters the first episode for a respondent, then each episode for that respondent would be examined in turn from the first to the last. If an episode has an activity code of 500, 510, 520, or 560 and an end time after 6:00 PM, 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 WGHTFIN for that respondent would be added to the estimate. In this example the sum over the 10,749 last episodes on the Time Use Episode File of WGHTFIN, when the derived variable is >Yes=, is 199,343 persons.

Estimates of the total time spent on some activities or groups of activities can be made either from the Main File or from the Time Use Episode File. For instance, to estimate the total amount of time Canadians over the age of 15 spend travelling for school and education in a day, one could form the weighted sum of the variable DUR590 on the Main File, or the weighted sum of the variable DURATION for all episodes with an activity code of >590' on the Time Use Episode File.

10,749

$$\sum_{i=1}^{10,749} \text{WGHTFIN} * \text{DUR590} = 77,958,520 \text{ minutes}$$

221,105

$$\sum_{i=1}^{221,105} \text{WGHTPEPI} * \text{DURATION} = 77,958,520 \text{ minutes}$$

(when ACTCODE=590)

In the first sum we are summing over all persons the total amount of time each spent travelling for school and education (a person based estimate), while in the second we are summing the episode duration over all episodes of travel for school or education (an episode based estimate).

An example of a total time spent estimate that could not be made from the Main File would be the total time spent on education in the evenings, which would be the weight sum over the Time Use Episode File of the duration after 6:00 PM of each episode with an ACTCODE between >500' and >590'.

221,105

$$\sum_{i=1}^{221,105} \text{WGHTPEPI} * \text{DURATION after 6:00 PM} = 126,131,637 \text{ minutes}$$

(when 500 <= ACTCODE <= 590)

Examples & Interpretation:

- (i) In 1998, 50.4% of female (SEX = 2) Canadians 15 years of age and older (6.2 million) stated they felt more rushed than compared to five years ago (A3 = 1).
- (ii) 56% of Canadians 25 to 44 years of age (2# AGEGR10 # 3) tend to cut back on their sleep, when they need more time for other activities (D2C = 1).
- (iii) 70% of males (SEX = 1) aged 15 to 24 (AGEGR10 = 1) stated that during the past 12 months they

regularly participated in sports (J1 = 1) while only 46% of females (SEX = 2) in the same age category took part regularly.

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 15 or more. When the number of contributors to the weighted estimate is less than this, the weighted estimate should not be released regardless of the value of the Approximate Coefficient of Variation.

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 on the questionnaire and errors may be introduced in the processing and tabulation of the data. These are all examples of non-sampling errors.

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

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 the interviewer was either unable to contact the respondent, no member of the household was able to provide the information, 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, 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 deviation 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.	Moderate Sampling Variability	0.0 to 16.5%	Estimates can be considered for general unrestricted release. No special notation is required.
2.	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.	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 policy 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 sampling variabilities for 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 estimates for Cycle 12 of the General Social Survey are based on a complex sample design, 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 result 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 below.

		Design Effects	
		<u>Geographic Area</u>	
		Canada	1.58
		Newfoundland	1.29
		Prince Edward Island	1.31
		Nova Scotia	1.29
Quebec	1.27	New Brunswick	1.29
		Ontario	1.28
		Manitoba	1.29
		Saskatchewan	1.29
		Alberta	1.23
		British Columbia	1.37
		Atlantic Region	1.36
		Prairie Region	1.33

Approximate variance tables for estimates using WGHTFIN are provided at the Canada and provincial levels as well as for the Atlantic and the Prairie Regions.

It should be noted that all coefficients of variation in these table 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 **A Highly Qualified** 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 WGHTFIN, and then using each of the bootstrap weights in place of WGHTFIN 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 is not releasable, then the quantitative total will in general not be releasable.

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 8 and the number 8500 rounded to thousands would be 9.

9. FILE STRUCTURE

In view of the nature of the time use data, the microdata file consists of the two subfiles described below.

The **Main File** consists of one record per respondent. It contains well over three hundred questionnaire-based variables. In addition it summarizes the total time spent on the diary day on each category of activity, the 10 major categories, the 24 subcategories¹, total time spent at each location, total time spent with various persons, and time spent helping persons outside the household.

This is the most widely used file for time use analysis. There are 10,749 records.

The **Time Use Episode File** consists of all episodes reported by respondents. Each respondent generated a variable number of records depending on the number of episodes reported. For each episode, there is information on the activity, start and end time, duration, location and an indication of who the respondent was with for that episode. New for Cycle 12, information on who an activity helped is included for selected types of activities. There are 221,105 records.

There is minimal duplication across the two files. However, the variable CASEID can be used for linking respondent characteristics and other main file variables with the detailed diary information on the Episode file. See Appendix N for guidelines for using Time Use data.

Special Notes

- 1 The diary response determined whether or not a respondent was included in the file. Therefore, the respondent may have stopped responding at any point from Section C of the questionnaire onwards. This methodological change may increase the rate of non-response to many questions and should be taken into account in any comparisons over time.
- 2 The sample and population counts for each variable in the data dictionaries are calculated from all respondents not only the ones specified in the coverage component of the description of the variable.

3. Not Stated Categories - Generally a code 9 for a one digit field, a code 99 for a 2 digit field, etc. indicate that the respondent did not answer a question and therefore the answer is not stated. As the following example indicates, two types of "Not Stated" categories may appear.

PLACE ²	Where were you?/Were you still...
01	Respondent's home
02	Respondent's work place
03	Someone else's home
04	Other place
05	Car (Driver)
06	Car (Passenger)
07	Walking
08	Bus and subway
09	Bicycle
10	Other form of transit
97	Missing or refused location (activity code is 001 or 002)
98	Unknown location
99	Not stated

Code 9, 99, etc. is the "true" not stated category for all variables on the file.

In certain questions, however, a second 'Not Stated' category appears. Although the respondent may not have marked a response, the information was actually partially available. Because of the branching pattern of a particular response, related information which followed allows imputations of the original question. Other responses within the question were truly not stated. These cases are thus identified separately.

10. ADDITIONAL INFORMATION

Additional information about this survey can be obtained from the individuals listed below. Data from the survey are available through published reports, special request tabulations, and this microdata file. The microdata file is available from the Housing, Family and Social Statistics Division of Statistics Canada at a cost of \$1,600.00. Tabulations can be obtained at a cost that will reflect the resources required to produce the tabulation.

Sample Selection Procedures, Weighting and Estimation

David Paton

Development and Analysis Section

Informatics and Methodology Field

(613) 951-1467

Subject Matter, Data Collection and Data Processing

Manon Declos

General Social Survey

Housing, Family and Social Statistics Division

(613) 951-9298

-
1. See Appendix L for the detailed structure.
 2. Variable is found in the Time Use Episode File. Part (d) of a diary episode asks the respondent where the activity they reported took place.