Income Statistics Division

62F0026MIE - 01001

1998 Survey of Household Spending Data Quality Indicators

Prepared by:
Sophie Arsenault
Jean-Luc Bernier
Jean-Marc Fillion
Sunita Patel
Johanne Tremblay

October 2001

Statistics Canada Statistique Canada

Canada

Data in many forms

Statistics Canada disseminates data in a variety of forms. In addition to publications, both standard and special tabulations are offered. Data are available on the Internet, compact disc, diskette, computer printouts, microfiche and microfilm, and magnetic tape. Maps and other geographic reference materials are available for some types of data. Direct online access to aggregated information is possible through CANSIM, Statistics Canada's machine-readable database and retrieval system.

How to obtain more information

Inquiries about this product and related statistics or services should be directed to: Client Services, Income Statistics Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 ((613) 951-7355; (888) 297-7355; income@statcan.ca) or to the Statistics Canada Regional Reference Centre in:

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(403) 495-3027
Ottawa	(613) 951-8116	Calgary	(403) 292-6717
Toronto	(416) 973-6586	Vancouver	(604) 666-3691
Winnipeg	(204) 983-4020		

You can also visit our World Wide Web site: http://www.statcan.ca

Toll-free access is provided **for all users who reside outside the local dialing area** of any of the Regional Reference Centres.

National enquiries line	**1 800 263-1136**
National telecommunications device for the hearing impaired	**1 800 363-7629**
Order-only line (Canada and United States)	**1 800 267-6677**

Ordering/Subscription information

All prices exclude sales tax

Catalogue no.62F0026MIE-01001, is available on internet for free. Users can obtain single issues at: http://www.statcan.ca/cgi-bin/downpub/research.cgi.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact your nearest Statistics Canada Regional Reference Centre.

Statistics Canada
Income Statistics Division

1998 Survey of Household Spending Data Quality Indicators

Published by authority of the Minister responsible for Statistics Canada

October 2001

Catalogue no. 62F0026MIE - 01001

Frequency: Irregular

Ottawa

La version française de cette publication est disponible sur demande

Note of appreciation

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

Abstract

This report describes the quality indicators produced for the 1998 Survey of Household Spending. It covers the usual quality indicators that help users interpret data, such as coefficients of variation, nonresponse rates, imputation rates and the impact of imputed data on the estimates. Added to these are various less often used indicators such as slippage rates and measures of the representativity of the sample for particular characteristics that are useful for evaluating the survey methodology.

The authors wish to thank the team responsible for the Survey of Household Spending in the Income Statistics Division (ISD) and also Sylvana Beaulieu and José Gaudet for their co-operation in the preparation of this report.

TABLE OF CONTENTS

TABLE OF CONTENTS (concluded)

HIGHLIGHTS

Sampling Errors

- The coefficients of variation (CVs) of the estimates of the average expenditure per household vary between 1.5% and 2.6% according to the province. The CVs are highest in the territories, namely 5.2% in the Yukon, 5.4% in the Northwest Territories and 8.9% in Nunavut.

Nonresponse

- The nonresponse rate is 23.6%. We were unable to contact nearly 5% of households. 16% of households refused to respond. Records for 3% of households were considered unusable because they were incomplete (1.6%) or were rejected because the differences between receipts and disbursements reported on the questionnaire exceeded 20% (1.3%).

- Analysis of response rates in the strata consisting of high and low-income geographic areas created under the sample design indicates that the nonresponse rates in high-income strata are approximately 65% greater than the rates observed in regular strata; there was a much higher frequency of refusals and incomplete data in these areas.

- Low-income households have a final nonresponse rate approximately 6 percentage points lower than households in regular strata and that result comes almost entirely from the result of different behaviour during collection since there is almost 5 percentage points less nonresponse for low-income strata compared to regular strata.

Coverage Errors

- The slippage rate for individuals 15 years of age and over is 9.6% indicating an undercoverage of the population, a rate similar to that observed in the Labour Force Survey (LFS) for the same period. The slippage rate for children (0 to 14 years of age) is 4.9% signalling overcoverage of children.

- The survey weights are adjusted to take slippage into account, but the bias will be small provided that the characteristics of the individuals omitted are similar to those of the individuals included in the provincial age groups used for adjustment.

Response Errors

- Response errors include recall errors, telescopic error and errors due to proxy response. Because the SHS interview is lengthy, the response burden can lead to respondent fatigue and have an impact on the data quality. For households providing usable data, the average time to complete the interview is 1 hour 50 minutes. Total interview time varies according to the household characteristics; for some households the interview can take more than five hours.

Processing errors related to imputation

i) Expenditure Variables

- 5.9% of respondents required some expenditure imputation with the majority of them having only one or two fields imputed out of the 237 expenditure variables (excluding the clothing section).

- The impact of imputed values on the total expenditure estimates is only 0.4% and is less than 1% for the estimate of the total for each summary level expenditure category excluding *Miscellaneous Expenses* (1,3%).

- Slightly more than 15% of the individuals required imputation for clothing variables. For the majority of these, the respondents provided the totals and the components were imputed.

ii) Income Variables

- Less than 3% of individuals required imputation for at least one income variable. For a little more than than 60% of these, total income was provided by the respondent and imputation was performed to get the breakdown by components.

- Imputed values contributed only 0.2% of the total income estimate.

INTRODUCTION

The Survey of Household Spending (SHS) is an annual survey that collects data on household income and expenditure using personal interviews conducted on a sample of approximately 24,000 households distributed throughout the ten provinces and two territories.[1] Collection takes place in January, February and March, and income and spending figures are obtained for the period from January 1 to December 31 of the previous year. This survey replaces the periodic Family Expenditure Survey. The main differences between it and the latter survey are that the level of detail in the questionnaire has been reduced, the sample has been increased to meet the need for provincial estimates by the Project to Improve Provincial Economic Statistics (PIPES), and the survey is now annual. The survey also becomes the collection vehicle for dwelling characteristics and household facilities and equipment replacing the Household Facilities and Equipment Survey. The 1997 survey, which refers to expenditures made in the year 1997, is the first to be conducted following this redesign.

Like all surveys, the SHS is subject to errors, despite all the precautions taken at the different stages of the survey to control them. While there is no comprehensive measure of the quality of the data generated by a survey, some quality measures produced at the different stages of the survey can provide users with the information needed in order to interpret the data properly.

This report therefore seeks to describe the quality indicators produced for the 1998 Survey of Household Spending. It covers the usual quality indicators that generally help users interpret data, such as coefficients of variation, nonresponse rates, imputation rates and the impact of imputed data on the estimates. Added to these are various less often used indicators such as slippage rates and measures of the representativity of the sample for particular characteristics that are useful for evaluating the survey methodology. Also included are the results of a few evaluations, such as an assessment of the models provided to users for calculating an approximation of CVs and an evaluation of the suppression rule that is used to determine whether an estimate is sufficiently reliable to be published.

Quality indicators have been classified according to the main types of error encountered in a survey. Section 1 deals with sampling errors—that is, errors due to the fact that the inferences about the population drawn from the survey are based on information collected from a sample of the population, rather than the entire population. The following sections cover errors not due to sampling. Nonresponse and coverage errors are first discussed in sections 2 and 3. Lastly, response errors and processing errors are dealt with in sections 4 and 5 respectively.

This report focuses on the data quality. For a detailed description of the methodology of the survey, see reference [6]. It should also be noted that a number of tables that are more detailed than those presented in this report may

[1] 1998 is the first survey year of SHS where we calculate estimates for Nunavut. In the previous years, Nunavut was part of the Northwest Territories estimates.

be obtained from the Household Survey Methods Division. The list of these tables is provided in Appendix A.

1. SAMPLING ERRORS

Sampling errors exist when inferences drawn from the survey about the population are based on information collected from a sample, rather than the entire population. In addition to the sample design and the estimation method used in the Survey of Household Spending, the sample size and the variability of each characteristic are factors that determine sampling error. Characteristics that are rare or are distributed very unevenly in the population will have greater sampling error than characteristics that are observed more frequently or are more homogeneous in the population.

1.1 Measures of Sampling Error

The standard error is a commonly used measure of sampling error. The standard error is the degree of variation of the estimate considering that a particular sample was selected, rather than another among all possible samples of the same size under the same sample design. Since the SHS uses a complex sample design and estimation method, the standard error is estimated using a resampling method known as the jackknife technique. For more details on this method, see reference [1] and [7].

The coefficient of variation (CV) is also a frequently used measure of the reliability of the estimate. It merely expresses the standard error as a percentage of the estimate. Thus, if an estimate Y is obtained for a certain characteristic and SE is the estimated standard error, then the CV will be (SE/Y) x 100.

Finally, either the standard error or the coefficient of variation may be used to derive another measure of the accuracy of estimates, namely the confidence interval. This measure indicates the level of confidence with which it can be stated that for a characteristic observed the real population or parameter value lies within the interval. An interval with a confidence level of 95% corresponds to the estimate obtained from the sample $\pm$ 2 standard errors (Y $\pm$ 2 SE).[2] This means that if the sampling were repeated a large number of times, each sample would provide a different interval and 95% of the intervals would contain the true value of the characteristic. Similarly, if the sampling were repeated, the interval Y $\pm$ SE would contain the true value in 68% of cases.

1.2 Coefficients of Variation

Estimates of coefficients of variation are calculated for estimates of many characteristics collected in the SHS. For a given expenditure characteristic, a number of estimates are produced, such as total household expenditure, average household expenditure, number of households reporting a value greater than 0, or average expenditure of households reporting a value greater than 0. Generally in the SHS, the CV of the estimate of average household expenditure is

[2] The confidence interval is calculated directly from the CV in similar fashion, namely Y $\pm$ 2 (CV x Y)/100.

published. For the 1998 SHS, this CV is available in the publication *Spending Patterns in Canada* (see reference [8]), at the national level for each of the detailed expenditures collected, as well as at the provincial or territorial level for several categories of summary level expenditures.[3] The CVs of detailed expenditure averages at the provincial or territorial level are available upon request from the Income Statistics Division. The CVs of dwelling characteristics and household facilities and equipment can also be obtained.

It should be noted that the estimated CVs do not consider the fact that some of the data were imputed and thus may underestimate the true CV's. For most variables, the impact of imputation is negligible (see section 6) and the provided CVs represent good estimates. For reliability of detailed expenditure with a high imputation rate, the CV and the impact of imputed data on the estimate should be considered simultaneously to make an assessment of the reliability.

Table 1.1 gives an overview of the CVs of estimates of household averages at the provincial and territorial level as well as at the national level for the estimation of a few of the summary level expenditure categories and for income.

[3] In previous surveys, CVs were published at the national level and for different income groups.

Table 1.1
Coefficients of Variation (%) by Province and Territory and for Canada for the Estimation of Average Household Expenditures for Several Summary Level Expenditure Categories and for the Estimation of Average Income

Summary level expenditure categories	*Can*	Nfld	P.E.I.	N.S.	N.B.	Que.	Ont.	Man.	Sask.	Alta	B.C.	Y.T.	N.W.T.	Nvt.
Total expenditure	0.8	2.5	2.6	2.1	1.9	1.6	1.5	1.8	1.8	1.6	2.0	5.2	5.4	8.9
Total current consumption	0.6	1.9	2.0	1.7	1.5	1.3	1.1	1.4	1.5	1.4	1.7	4.3	3.6	8.0
Food	0.6	1.3	1.5	1.3	1.7	1.1	1.1	1.2	1.3	1.2	1.3	4.1	4.4	7.9
Shelter	0.8	2.9	2.9	2.1	2.5	1.4	1.5	1.7	2.1	1.8	1.8	3.0	7.4	17.2
Household operation	1.0	2.4	2.8	2.6	2.4	1.9	1.9	1.8	2.2	1.9	2.4	5.7	5.9	10.4
Furnishings	2.2	3.6	4.2	3.5	3.5	3.5	4.2	3.8	3.7	3.8	6.2	7.6	10.1	10.6
Clothing	1.1	2.6	3.2	2.9	2.6	2.1	2.1	2.5	2.5	3.5	2.5	7.7	6.3	4.5
Transportation	1.4	4.1	4.4	3.6	3.3	2.9	2.6	3.7	3.2	3.2	4.0	9.1	5.0	30.6
Health care	1.3	3.7	3.9	3.5	2.6	2.2	3.2	3.3	3.0	2.2	3.0	7.0	5.7	19.3
Personal care	1.0	2.4	3.0	2.8	2.1	2.0	1.9	2.7	2.3	2.1	2.4	5.6	6.4	12.3
Recreation	1.6	3.7	6.1	3.6	3.8	4.0	2.7	3.6	3.5	3.6	3.9	8.9	9.1	9.1
Reading & printed materials	1.4	4.4	4.2	3.6	3.1	2.9	2.7	3.9	3.0	2.9	2.9	12.3	5.5	19.7
Education	3.3	8.5	13.7	11.6	10.3	6.1	6.6	8.2	7.7	5.7	7.1	24.3	9.9	42.2
Tobacco, alcoholic beverages	1.5	4.2	5.7	5.1	3.6	2.5	3.2	4.0	4.4	4.0	4.0	7.1	9.5	25.1
Games of chance (net)	3.2	6.0	8.6	8.3	8.2	4.9	7.0	8.7	11.3	8.6	6.5	17.3	23.5	15.0
Miscellaneous expenditures	3.2	7.1	7.7	7.4	6.2	8.2	6.0	10.3	6.5	4.1	6.7	10.0	9.3	8.7
Personal income tax	1.8	5.0	5.2	4.2	4.0	3.1	3.6	4.2	4.0	3.0	4.3	9.2	8.6	17.0
Personal insurance	1.5	5.3	4.7	3.6	3.6	3.1	3.0	5.2	7.9	2.3	3.1	8.9	7.0	11.3
Gifts and contributions	3.3	7.0	16.9	8.0	7.4	8.1	6.5	7.9	7.4	6.5	7.5	30.2	44.2	33.8
Income	1.0	2.6	2.6	2.2	2.1	1.8	1.9	2.0	2.1	1.6	2.0	6.2	5.0	8.1

The coefficients of variation of the estimate of total expenditures per household vary between 1.5% and 2.6% at the provincial level. The CVs are higher in the territories, namely 5.2% in the Yukon, 5.4% in the Northwest Territories and 8.9% for Nunavut.

For the different categories of summary level expenditures, the CVs at the national level are also less than or equal to 2.2%, except for expenditures in the following categories: *education*, *games of chance*, *miscellaneous expenditures* and *gifts and contributions*. These expenditure categories account for respectively 1.3%, 0.5%, 1.6% and 2.2 % of the total expenditure.

Table 1.2 gives an overview of the CVs for some dwelling characteristics and household equipment estimates at the provincial and territorial level as well as at the national level.

Table 1.2
Coefficients of Variation (%) at the National, Provincial and Territorial Level for some Characteristics of Household Facilities and Equipment

CATEGORIES	*Can*	Nfld	P.E.I.	N.S.	N.B.	Que.	Ont.	Man.	Sask.	Alta	B.C.	Y.T.	N.W.T.	Nvt.
Owner	1.0	1.9	3.3	2.1	2.3	1.8	2.1	2.3	1.9	2.0	1.9	8.3	13.8	5.6
Renter	1.8	6.2	7.7	5.6	6.3	2.4	4.2	5.1	4.3	4.4	3.4	15.4	12.4	1.7
Automatic washing machine	0.6	1.4	2.1	1.7	2.0	0.9	1.5	1.7	1.5	1.2	1.5	6.8	6.3	10.2
Clothes dryer	0.6	1.2	2.3	1.7	2.0	1.0	1.5	1.7	1.4	1.2	1.5	6.8	7.7	9.2
Built-in dishwasher	1.5	5.9	6.0	4.8	5.1	2.5	3.3	3.8	3.8	2.8	2.9	11.5	11.7	27.0
Freezer	0.9	1.6	3.0	2.2	2.3	2.2	1.9	1.8	1.4	1.8	2.1	8.1	7.0	13.4
Microwave oven	0.4	1.1	1.5	1.0	0.9	0.8	0.8	1.0	1.0	0.7	1.0	2.3	3.8	9.2
Cellular phone	2.0	7.8	9.6	6.1	6.4	5.4	3.8	5.1	4.5	2.9	3.6	16.4	21.0	-
CD player	0.7	2.0	2.7	2.2	2.1	1.5	1.5	2.0	2.0	1.4	1.6	4.2	2.9	7.7
Cable TV	1.1	2.0	4.2	2.0	2.3	1.8	2.5	1.7	2.7	1.8	1.5	11.7	4.2	10.4
Home computer	1.3	4.0	6.0	3.8	4.7	2.9	2.4	3.0	3.4	2.4	2.4	5.7	5.4	11.5
Modem	1.8	4.8	8.0	5.5	5.9	4.1	3.4	4.0	4.6	3.2	3.3	6.9	10.2	22.5
Use of internet (home)	2.1	5.9	9.5	6.6	6.5	4.9	4.0	4.8	5.7	3.8	4.3	8.4	10.9	28.9
Owned vehicles (one)	1.3	3.8	4.8	3.5	3.9	2.3	2.9	2.8	3.0	3.0	2.9	8.9	11.1	25.7
Owned vehicles (2 or more)	1.5	4.8	4.7	4.0	3.8	3.4	3.1	3.2	2.9	2.5	2.9	10.4	20.2	60.0

The coefficients of variation for the estimates of the characteristics of Household Facilities and Equipment are generally below 6% for each province and variable except the following categories: renter, cellular phone, modem and use of internet. The CV's are higher in the territories where we can find a smaller proportion of equipment.

The coefficients of variation for the characteristics of household facilities and equipment at the national level are below 2.1%.

1.3 Model for Deriving an Approximation of the CV

For operational reasons, it is not possible to produce CVs for all the characteristics collected by the survey at all the different levels of aggregation that may interest users, such as by income quintile, household type, urbanization level or tenure, or for selected metropolitan areas. However, such estimates are available at these different levels for the summary level expenditure categories in the publication *Spending Pattern in Canada* (see reference [8]), and they may be obtained for detailed expenditures from the Income Statistics Division.

1.3.1 Model for Deriving an Approximation of the CV for Domain Estimates

It is also possible to calculate an approximation of the CV by using a relationship between the number of households in the sample that reported expenditures for a given category and the CV at an aggregated level. This relationship, based on the CV's tendency to increase in proportion to a decrease in the square root of the number of households reporting an expenditure, is illustrated on the following page.

Formula for Approximating the CV for a Domain (Subgroup of the Population)

If CV(Y) represents the CV for the estimate of the average per household of a certain characteristic for the entire population, then an approximation of the CV of the estimate of that characteristic can be calculated for a domain (which may be considered as a subgroup of the population, such as household type, an income quintile, an urbanisation level, etc.) according to the following equation:

$$CV(Y_d) = CV(Y) \times \sqrt{\frac{nP}{n_d P_d}}$$

where

n: *number of households in the sample*
P: *estimate of the proportion of households reporting a value > 0 for this characteristic in the population*
n_d: *number of households in the sample in domain d*
P_d: *estimate of the proportion of households reporting a value > 0 for this characteristic in domain d*

Generally, approximations for the different domains are calculated using the CV, size *n* and proportion *P* at the national level. If an approximation of the CV for a metropolitan area is desired, these values can be used at the provincial level, since the domain is entirely contained within a single province and the provincial CVs are published for the 1998 SHS.

1.3.2 Method of Computation of an Approximate CV from the Microdata File

The microdata file users can obtain an approximation of the CV of the estimates from another method which will generally provides better results than the method described in the previous section for the CVs of detailed expenditure estimates. This approach is described in detail in the documentation provided with the 1997 microdata file. This method of approximation can be used only for the microdata file since it is necessary to have data and weights for each household.

The 1997 data quality document of the survey contains the results of the evaluation of the performance of the two CV approximation methods that we have just described.

1.4 Suppression of Unreliable Data in Estimation Tables

Since the coefficient of variation is an indicator of the reliability of the data, we would like to use it to determine whether or not the estimates should be published. Estimates for which the CV is more than 33% are not considered sufficiently reliable to be published. However, CV estimates are not calculated for many of the published estimates. The suppression rule for expenditure estimates is therefore based on the number of households reporting a value greater than 0.[4]

It can be shown that CVs generally reach roughly 33% when the number of households reporting an expenditure approaches 30. Since this is an approximate rule, some estimates may be published even though the CV is greater than 33%, and some estimates will not be published even though the CV is less than 33%.

Since the summary CV calculated for the evaluation of the model in section 1.3 include only a few cases where the number of households declaring an expense below 30, only the data at the detailed level were used for estimating the risk of error when suppressing. The document on data quality for the 1997 SHS give the results of the evaluation of the risk of error in the use of the suppression rule.

2. NONRESPONSE

Errors due to nonresponse result from the fact that some potential respondents do not provide the necessary information or the information proves to be unusable. Where the respondent has failed to respond to only some questions, this is referred to as partial nonresponse. In such a case, the missing data are imputed. Errors associated with imputation are described in Section 5, which deals with processing errors. In the present section, nonresponse includes collection nonresponse, which is mainly due to inability to contact the household or the refusal of the members of the household to participate partially or completely in the survey, as well as data collected from households that prove to be unusable.

The main impact of nonresponse on data quality is that it can introduce a bias in the estimates if the characteristics of respondents and nonrespondents differ and this difference has an impact on the characteristics studied. Nonresponse rates may easily be calculated, but they have only an indicative value with regard to data quality, since they do not allow estimation of the bias associated with the estimates. The scope of nonresponse may be considered as an indicator of the risks of bias in the estimates.

2.1 Nonresponse Rates and Vacancy Rates

In the SHS, since the units selected are dwellings, interviewers must first identify ineligible dwellings—that is, dwellings occupied by persons who are not part of the target population—as well as dwellings that no longer exist (demolished,

[4] In practice, we use the estimate of the proportion of households reporting an expenditure, which is multiplied by the sample size.

mobile home moved or dwelling converted to business location) and vacant dwellings (unoccupied, seasonal or under construction).

Among eligible dwellings, we next evaluate the proportion of households that did not respond to the survey, which is called the collection nonresponse rate. These include households that refused to participate in the survey and households where no contact could be made with the respondents, either because they were absent or because of special circumstances (language problem, illness, death).

Again among eligible dwellings, the rate of unusable data is determined. Unusable data refers to the number of households whose questionnaire was at least partially completed but which were rejected during the processing of the data. There are two main causes of rejection. First, when many of the questions on income or the questions on expenditures have been left unanswered, the questionnaire is classified as incomplete and is not used. The other source of rejection consists of questionnaires in which the difference between receipts (income and other sources of money received by the household) and disbursements (expenditures and net change in assets and liabilities) is greater than 20%. These questionnaires are also excluded from the estimation and are considered as nonresponse.

Table 2.1 shows the final nonresponse rate as well as the collection nonresponse rate, broken down by refusals and no-contacts, and the rate of unusable data broken down into incomplete and unbalanced questionnaires. The vacancy rate is also included. These rates are provided at the national level as well as at the provincial or territorial level.

Table 2.1
Nonresponse Rate (%) and Vacancy Rate (%) by Province or Territory

Province or territory	Vacancy rate	Collection nonresponse rate			Unusable data rate			Final nonresponse rate (at estimation stage)
		TOTAL	No contact	Refusal	TOTAL	Incom-plete	Un-balanced	
Canada	**10.5**	**20.7**	**4.9**	**15.8**	**2.9**	**1.6**	**1.3**	**23.6**
Newfoundland	15.1	14.1	4.2	9.9	2.8	1.5	1.3	16.8
P.E.I.	17.4	16.5	2.0	14.5	1.5	1.2	0.2	18.0
N.S.	15.0	18.9	4.6	14.3	5.9	2.7	3.2	24.8
N.B.	11.2	14.6	3.0	11.6	1.6	0.9	0.7	16.2
Quebec	8.6	22.8	4.7	18.1	0.6	0.4	0.2	23.5
Ontario	7.1	26.3	6.3	20.0	4.2	2.4	1.8	30.5
Manitoba	12.5	18.2	3.3	14.9	1.8	1.5	0.3	20.0
Sask.	10.9	14.9	4.2	10.7	1.8	1.6	0.2	16.7
Alberta	6.6	22.1	5.2	17.0	1.7	1.3	0.4	23.8
B.C.	7.2	28.2	7.5	20.7	5.5	2.0	3.6	33.7
Yukon	17.2	20.6	6.8	13.8	7.3	3.7	3.7	27.9
N.W.T.	16.2	10.7	1.8	8.9	0.8	0.5	0.3	11.5
Nvnavut	13.1	3.5	2.0	1.5	2.0	2.0	0.0	5.5

The final nonresponse rate at the national level is 23.6%. It is mainly due to refusals (16%), households that we were unable to contact (5%), and finally to households for which the data were unusable (3%). Households with unusable data break down almost equally into those with incomplete data and those with unbalanced questionnaires.

The final nonresponse rate varies a lot from one province to another. Nunavut and the Northwest Territories register the lowest rates at 5.5% and 11.5% respectively, while Newfoundland, Prince Edward Island, New Brunswick, Manitoba and Saskatchewan range between 16% and 20%, and rates in excess of 30% are observed for Ontario and British Columbia. Both the refusal rates and the no contact rates of the latter two provinces are higher than those of the other provinces. The unusable data rate varies greatly from one province to another. It is especially low in Quebec as well as in the Northwest Territories at less than 1%. The rate is higher than 5% in Nova Scotia, British Columbia and the Yukon.

The vacancy rate is shown in Table 2.1, but it should be kept in mind that vacant dwellings do not contribute to the bias of the sample if they are correctly identified. By analysing vacancy rates, we can detect dwelling identification problems associated with the collection process. The national vacancy rate for the 1998 SHS is 10.5%, which is slightly lower than for the rate for the previous year, which was around 11%. Note that the vacancy rate for the 1998 SHS is slightly lower than the rate for the Labour Force Survey (LFS) for the same period.

2.2 Nonresponse According to Urbanization Level

Nonresponse varies according to urbanization level. The various rates at the national level are shown by urbanization level in Table 2.2.

Table 2.2
Nonresponse Rate (%) and Vacancy Rate (%) by Urbanization Level

Urbanization category	Vacancy rate	Collection nonresponse rate			Unusable data rate			Total nonresponse rate) (at estimation stage)
		TOTAL	No contact	Refusal	TOTAL	Incom-plete	Un-balanced	
URBAN								
1,000,000 or more	4.1	28.6	7.2	21.3	3.2	1.5	1.7	31.7
500,000 to 999,999	3.8	22.9	4.2	18.7	1.4	1.1	0.3	24.3
250,000 to 499,999	5.4	26.3	6.8	19.5	5.7	2.5	3.1	32.0
100,000 to 249,999	7.0	19.8	5.2	14.6	2.7	1.7	1.0	22.5
30,000 to 99,999	5.9	18.5	4.3	14.2	3.0	1.5	1.5	21.5
Less than 30,000	10.7	16.2	3.8	12.4	2.7	1.8	0.9	18.9
RURAL	22.4	15.1	3.4	11.6	3.1	1.6	1.4	18.2
TOTAL	**10.5**	**20.7**	**4.9**	**15.8**	**2.9**	**1.6**	**1.3**	**23.6**

Generally speaking, the nonresponse rate increases with the level of urbanization. According to table 2.2, only the "500,000 – 999,999" group goes against this pattern with a final nonresponse rate of 24%.

For the collection nonresponse rate, there is a notable difference of at least 3% between the urbanization levels of 250,000 inhabitants and more and the other urbanization levels. Refusals account for more than 60% of the total nonresponse at each level of urbanization.

From an examination of the vacancy rate by urbanization level, it is clear that the vacancy rate is much higher in rural areas (22%) and low-population urban areas of less than 30,000 (11%) than in high-population urban areas. This phenomenon is also observed in the LFS and may be explained by a greater number of seasonal dwellings in rural areas. This also explains the higher vacancy rate in the Atlantic provinces, as illustrated in Table 2.1, and especially in Prince Edward Island, which has a high proportion of rural dwellings. Since the SHS sample is more concentrated in high-population urban areas than the LFS, the national vacancy rate for the SHS can be expected to be slightly lower than that for the LFS.

2.3 Nonresponse According to Income Strata

Since income information is not available for nonrespondents, it is not possible to compare nonresponse rates according to income. However, the LFS sample design, used for the SHS, was designed in such a way that in seven large cities there are strata consisting of geographic areas where the average household income exceeds $100,000 as well as strata consisting of apartments inhabited by households with an average income of less than $20,000. Even though the

number of such strata is small and accounts for only a small number of dwellings in the SHS sample (approximately 280 and 200 for high incomes and low incomes respectively, or less than 2% of the sample), the comparison of nonresponse rates in these two groups in relation to the other strata is revealing. Table 2.3 shows these results.

Table 2.3
Comparison of Nonresponse and Vacancy Rates (%) in High-income and Low-income Strata in Relation to Other Strata

Stratum type based on income	Vacancy rate	Collection nonresponse rate			Unusable data rate			Final nonresponse rate (at estimation stage)
		TOTAL	No contact	Refusal	TOTAL	Incom-plete	Un-balanced	
High-income	3.6	35.3	8.1	27.2	3.0	2.1	0.9	38.3
Regular	10.6	20.6	4.8	15.7	2.9	1.6	1.3	23.5
Low-income	5.4	16.1	5.4	10.7	1.8	0.6	1.2	17.9
TOTAL	**10.5**	**20.7**	**4.9**	**15.8**	**2.9**	**1.6**	**1.3**	**23.6**

The final nonresponse rate in high-income strata is approximately 65% greater than the rate observed in the regular strata, at slightly more than 38%. The refusal rate for the high-income strata is just over one in four households (27%), a rate almost 75% higher than that for regular strata. Only the nonresponse caused by unbalanced data gives a rate for high-income strata that is less than the two other strata.

Households in low-income strata have a final nonresponse rate approximately 6 percentage points lower than households in regular strata resulting almost entirely from different behaviour during collection since the refusal rate for low-income strata is 5 percentage points less than that for regular strata.

As for SHS 1997, the vacancy rate is higher for regular strata than for each of the two other strata, even though the difference is smaller than for the 1997 survey.

2.4 Adjustment for Nonresponse

To compensate for nonresponse, the weights in the SHS are inflated by the inverse of the weighted response rate within certain groups defined on the basis of the different urbanization levels in each province or territory. The weighted rates differ from the rates presented in this section since the former take into account the sampling weight of each household. An algebraic description of the nonresponse adjustment is provided in Appendix B.

The adjustment of weights for nonresponse takes into account the differences in nonresponse by urbanization level as described in Section 2.2. It will reduce the bias to the extent that the characteristics of respondents and nonrespondents are similar for a given urbanization level.

3. COVERAGE ERRORS

In the design of the survey, the target population was defined. It is useful to go over this definition, since a good understanding of the target population is necessary in order to properly interpret the survey data. It is important to note that in the SHS, the target population is different for the provinces and territories.

Target population

> The target population consists of individuals living in private households. It therefore excludes residents of institutions such as prisons, chronic care hospitals or senior citizens' homes, as well as members of religious orders and other groups living communally, members of the Armed Forces living in military compounds, and individuals residing permanently in hotels or rooming houses. Also excluded are foreign countries' official representatives residing in Canada and their families as well as individuals residing on Indian reserves or public lands. With these exclusions, the survey covers nearly 98% of the population in the ten provinces. In the Yukon, persons living in small communities or in unorganized areas are also excluded, and the survey covers approximately 81% of the population. The coverage of the Northwest Territories represents 92% of the population and the coverage of the Nunavut represents 89% of the population.[5]

Coverage errors result from inadequate representation of the target population based on the units of the sampling frame. Some units of the target population may be omitted from the sampling frame, in which case there is undercoverage. Other units that are not in the target population may be included by error, or some units may be included more than once; these units are responsible for overcoverage.

3.1 Undercoverage and Overcoverage: Slippage Rates

In the SHS, the sample is selected using a list of dwellings in each selected cluster. Factors contributing to undercoverage are the omission of dwellings in the creation of the list, new dwellings that are added between the creation of the list and the interviewer's visit (mainly in developing areas) as well as the erroneous classification of vacant dwellings. The inclusion of dwellings that are not within the boundaries of the cluster is a source of overcoverage. Similarly, errors can take place due to improper identification of persons as members of the selected household during data collection. These errors also contribute to undercoverage or overcoverage.

A good representation of the target population is essential to the production of realistic expense estimates. The number of people per household is also an important characteristic in the estimation of household average expenses. Therefore, it is necessary that the sample not only adequately represents the

[5] In terms of households, the coverage of the Yukon, the Northwest Territories and the Nunavut represents respectively 80%, 93% and 90% of households.

individuals in the target population, but also the distribution of households according to their size. SHS 1998 data were reweighted using a strategy that was first introduced for the 1999 survey. This weighting strategy utilises new controls in order to better correct the representation of the target population.

There is generally net undercoverage of the number of persons in the SHS, which is corrected by an adjustment of weights using post-censal demographic estimates. The slippage rate (see appendix B) represents the percentage difference between survey estimates calculated using weights not adjusted with external data[6] and post-censal demographic estimates. A positive rate indicates overcoverage of the number of persons in the survey. Tables 3.1 and 3.2 respectively show slippage rates by age-sex group at the national level and at the provincial or territorial level, while Table 3.3 presents these rates for the household size categories used for the weight adjustment.

Table 3.1
National Slippage Rates by Age-Sex Group

National Slippage Rates (%) by Age-Sex Group				
	Age	Sex		**Total**
		Male	Female	
Canada	0-6 years	7.4	7.3	7.4
	7-17 years	6.7	0.4	3.6
	18-24 years	-12.0	-12.17	-12.1
	25-34 years	-11.5	-6.5	-9.0
	35-54 years	-8.0	-6.3	-7.1
	55-59 years	-13.9	-7.8	-10.8
	60-64 years	-10.7	-8.2	-9.4
	65-69 years	-9.7	-6.8	-8.2
	70 years and +	-7.7	-10.3	-9.3
	Total	**-5.7**	**-5.3**	**-5.5**

For the 1998 SHS, the national undercoverage rate was 5.5%. The rate varied from 1% to 12% in the provinces and reached a much higher level in the Yukon, the Northwest Territories and Nunavut (17%, 20% and 20% respectively). With respect to age group, we can see that national slippage rates for children (0 to 6 and 7 to 17) are quite different from those for other age groups. Children are over-represented while adults are always under-represented. The highest national rates occurred among 18 to 34 year old and 55 to 64 year old for the men, and among 18 to 24 year old and 70 year old and older for the women. For both sexes considered together, the highest undercoverage rates occurred for the 18-24 and the 55-59. It also appears that slippage rates generally differ by sex.

Provincially (Table 3.2), total undercoverage in New Brunswick, Quebec, Ontario and Manitoba was slightly less than the undercoverage observed nationally. The opposite was observed for the other provinces and for the territories. However, a

[6] The subweight which is the survey weight adjusted for nonresponse is used (see Appendix B)

low overall rate of undercoverage is no guarantee of better coverage. For example, Manitoba overall slippage rate (-1.2%) concealed the two worst cases of overcoverage for a provincial age-sex group (20.7 % among 17-24 women and 18.2 % among the 0-6 men). We can also see that the pattern of slippage rates differs substantially for age-sex groups from one province to the next.

Table 3.2
Slippage Rates for Provinces and Territories by Age-Sex Group

Slippage Rates (%) by Age-Sex Group														
Sex	Age	Newfoundland	Prince Edward Island	Nova Scotia	New Brunswick	Quebec	Ontario	Manitoba	Saskatchewan	Alberta	British Columbia	Yukon	Northwest Territories	Nunavut
Male	0-6	12.3	-27.9	4.1	4.9	9.8	13.3	4.4	4.1	-6.1	-0.1	*	-13.7	-15,5
	7-17	-13.8	-5.5	-6.6	14.2	7.8	11.6	20.7	0.5	3.9	-4.9			
	18-24	-27.0	-24.9	-25.8	-14.6	1.8	-14.0	-9.6	-23.4	-12.5	-23.0	-28.5	-24.4	-26.4
	25-34	-14.0	-24.2	-18.9	-21.5	-6.6	-10.9	-8.1	-11.6	-18.0	-13.6			
	35-54	-17.5	-5.0	-4.7	-0.4	-9.5	-4.4	-6.3	-12.7	-9.5	-14.5			
	55-59	2.3	-18.4	-19.7	-21.1	-16.5	-9.1	-7.5	-13.7	-31.8	-11.7			
	60-64	-6.6	19.8	4.2	3.6	-10.1	-12.3	-3.8	5.9	-9.2	-21.8			
	65-69	9.2	-12.4	-5.7	-11.3	-7.7	-15.5	14.2	4.5	3.2	-17.6			
	70 +	-22.6	-1.7	-2.6	-11.9	-1.4	-11.5	-5.9	-8.4	-9.9	-5.9			
	Total	-13.1	-11.6	-8.6	-4.4	-3.6	-3.6	-0.7	-8.1	-9.1	-12.2	*	-21.0	-21.7
Female	0-6	-4.3	-9.6	-8.6	14.5	3.1	13.2	18.2	5.9	-3.8	7.2	*	-12.7	-13.0
	7-17	-7.9	-10.8	-3.7	1.0	4.4	3.6	-3.5	1.3	-10.5	-4.0			
	18-24	-28.2	-28.3	-22.2	-12.8	-18.7	-4.5	-11.8	-5.4	-10.9	-18.9	-16.1	-20.3	-21.1
	25-34	-6.6	-6.5	-25.6	-2.1	0.8	-7.1	-9.2	-7.5	-6.7	-12.9			
	35-54	-14.9	-11.5	-1.8	0.3	-6.6	-4.4	0.3	-10.9	-9.9	-10.0			
	55-59	-1.6	16.8	-12.1	4.8	-7.7	-1.9	0.2	-8.0	-17.6	-23.7			
	60-64	5.8	7.8	-11.1	-8.9	-13.8	-4.0	-8.1	-2.3	-10.0	-11.1			
	65-69	6.2	-15.0	-7.5	2.7	-1.7	-6.1	1.1	-8.6	4.2	-30.8			
	70 +	-11.9	-4.2	-6.8	-5.6	-11.1	-17.5	-1.8	5.1	-8.7	1.0			
	Total	-10.8	-9.5	-9.5	-0.8	-5.0	-3.3	-1.7	-4.0	-8.8	-9.4	*	-17.8	-17.5
Total		**-11,9**	**-10.5**	**-9.1**	**-2.6**	**-4.3**	**-3.4**	**-1.2**	**-6.0**	**-9.0**	**-10.8**	**-17.1**	**-19.5**	**-19.7**

* The slippage rate for children aged 0-17 in the Yukon was -2.3%. There was no breakdown by sex for this age group in the Yukon.

Nationally, the number of households is underestimated by 5.1%. See table 3.3. This slippage rate is comparable to the -5.5% slippage rate for individuals. The one-person households are more under-represented than the others while the two-person households under-representation is slightly smaller than this observed for the household of size 3 and more.

Manitoba (0.1 %) and New Brunswick (-1.5 %) show the best coverage of households. For the one-person households, Yukon (-30.7 %), Ontario (-9.4 %)

and Quebec (-7.1%) are the only ones to with an undercoverage rate being higher than this observed nationally.

Table 3.3
Slippage Rates for Provinces and Territories by Household Size

Province or Territory	Slippage Rate (%)			
	Number of households	Number of one-person households	Number of two-person households	Number of three-person and plus households
Canada	**-5.1**	**-6.8**	**-4.0**	**-4.8**
Newfoundland	-8.4	-6.3	-1.7	-12.8
Prince Edward Island	-3.9	-1.2	7.6	-12.9
Nova Scotia	-5.6	-5.9	1.8	-11.1
New Brunswick	-1.5	-2.0	1.0	-3.0
Quebec	-4.7	-7.1	-4.0	-3.6
Ontario	-4.2	-9.4	-5.8	-0.6
Manitoba	0.1	0.9	2.5	-2.3
Saskatchewan	-3.8	-3.4	1.0	-7.9
Alberta	-6.3	-5.1	-3.1	-9.1
British Columbia	-8.9	-5.2	-5.9	-13.6
Yukon	-16.6	-30.7	-9.3	-13.1
Northwest Territories	-7.6	-1.6	10.2	-16.4
Nunavut	-32.4	-32.4[7]		

3.2 Adjustment for Coverage Error at the Population Level and Household Levels

To correct the coverage problem illustrated in tables 3.1 and 3.2 and reduce the resulting bias, the survey data are adjusted during weighting using demographic estimates for the age groups defined in the table, for each province or territory. For more information on the methodology of the adjustment, see reference[6]. This adjustment greatly reduce the bias caused by coverage errors but does not completely eliminate bias if the characteristics of the individuals omitted from the survey differ from those of the included individuals for a given age group in a province or territory.

Furthermore, the effectiveness of the coverage adjustment based on demographic estimates depends mostly on the quality of those estimates and their accuracy in representing the target population of the survey. The demographic estimates are not error-free. They are post-censal estimates based on the population counts from the 1996 Census adjusted for net undercoverage, and they take into account recent statistics on migration, births, deaths, etc. These demographic estimates are adjusted to account for certain exclusions specific to household surveys, such as persons living in institutions. Conceptually, they differ slightly from the SHS target population in that they include persons living in non-institutional collective dwellings, such as members

[7] Only the total number of households was used for Nunavut.

of groups living communally and individuals permanently residing in hotels or rooming houses. However, this difference is considered negligible, since such individuals represent less than 0.4% of the Canadian population.

To remedy the issue of the sample's representivity with respect to the number of households based on their respective sizes as illustrated in Table 3.3, we use supplementary data to adjust the data appearing in the survey. By adjusting the weight of the SHS to reflect post-census estimates of the number of households by size, we hope to compensate for the bias produced by inadequate representation of households. However, we will not necessarily succeed in eliminating such bias if features of uninterviewed (omitted or non-respondent) households differ from those of responding households for the same size or group. Naturally, the success of such an adjustment depends on the quality of the supplemental data.

In addition to demographic estimates of age-sex groups by province and territory, three other groups of supplementary data are used during weighting to adjust survey data and thereby improve their representativity. The first set of data is used to control for the number of children and adults in certain major cities. The second is designed to control for the number of single-parent households and couples with children by province. Finally, counts for major categories of income from wages and salary are used when adjusting weights to ensure a certain degree of consistency between the income distributions from the SHS and from outside sources.

4. RESPONSE ERRORS

Response errors represent a lack of accuracy in responses to questions. They can be attributed to different factors, including a questionnaire that requires improvements, misinterpretation of questions by interviewers or respondents, and errors in respondents' statements.

In the SHS, there can be various reasons for errors in respondents' statements. First, there are recall errors that occur when a respondent forgets expenditures made during the period covered by the survey (which corresponds to the calendar year) or provides an erroneous value because of the time interval that has elapsed between the time of purchase and the date of the interview. Recall errors are probably the survey's largest source of response error, since the reference period is long (12 months) and a great variety of information is requested.

One of the main measures taken to minimize recall error in the SHS is to calculate, for each household, the difference between receipts (income and other amounts received by the household) and disbursements (expenditures plus net change in assets and liabilities). When the difference exceeds 10% of receipts or disbursements, with the higher amount being retained, respondents are contacted again in order to obtain additional information and to try to identify errors or omissions. The respondent is also encouraged to consult various documents (invoices, bank statements, etc.) in order to provide more accurate data. To determine expenditures for small items purchased at regular intervals,

interviewers generally suggest to respondents that they estimate the frequency of the purchases and the price generally paid in order to derive expenditures for a twelve-month period.

A second source of error in respondents' reporting is telescopic error, which consists of including in the reference period events that occurred before it. In the SHS, the use of the calendar year is considered to provide a good marker for the start of the reference period. Furthermore, since the reference period is a long one, telescopic error has less impact.

Responses by proxy can also contribute to response error. The household member who made an expenditure is generally best able to report it accurately. This is definitely the case with, say, personal purchases. Expenditures reported by an intermediary are more likely to be tainted by response error, and this type of error tends to have a greater effect on certain types of expenditures.

Among other sources of response error, the extent of the respondent's co-operation should not be overlooked. For personal reasons, the respondent may decide not to mention particular expenditures or twist the facts.

In the SHS, another factor is the response burden, owing to the length of the interview and the great variety of items to be reported, as well as the pace of the interview. This can lead to respondent fatigue and affect the quality of the responses obtained. For respondent households that have supplied usable data, the average time needed to complete the interview is 1 hour and 50 minutes. The interview time varies greatly from one household to another, depending on household size, income and various other characteristics. For some households, the interview can take more than 5 hours.

While response errors are a major source of error in a historical interview, they are the aspects of data quality that are the hardest to measure. Generally, to attempt to measure them, it is necessary to conduct quite costly special studies. Efforts are made to combat response errors by using survey techniques designed to reduce them.

5. PROCESSING ERRORS

Processing errors can arise in all types of data handling. The main stages of data processing are coding, data entry, editing, imputation of partial nonresponse and weighting. In the SHS different procedures are applied at each stage in order to minimize processing errors and the survey estimates are compared with other data sources prior to release. Errors related to the adjustments made at the weighting stage have been described in sections 2 and 3. The other types of processing errors are covered in this section.

Coding is necessary for only a few questions. This is done by the interviewer and subsequently verified by a senior interviewer. Data entry is done with the help of an automated verification system that groups the questionnaires into batches and chooses some questionnaires from each batch to be entered a second time. Any errors found will then be corrected. If the number of errors in a

batch is greater than a certain threshold, then the entire batch is submitted for re-entry.

The first stage of automated verification is done after each questionnaire has been verified manually by both the interviewer and the senior interviewer. It is ensured that the respondent's answers respect some essential consistency rules. Unusual situations that may justify corrections are also identified. This stage of verification is done in the Statistics Canada regional offices in case it is necessary to recontact respondents if some supplementary information is required to resolve inconsistencies in their answers. Members of the verification teams that received special training in this area solve any problems identified. Thereafter, other verification checks are done at head office and invalid responses are corrected.

The processing of SHS data also involves imputation for partial nonresponse. Partial nonresponse occurs when the respondent refuses to answer or doesn't know the answer to certain questions. The imputation approach differs depending on whether the data is categorical or continuous. Categorical data takes on only specific values as in yes/no questions or type of dwelling questions, while continuous data can take any numerical value as for income and expenditure data.

Categorical data, which are obtained mainly in the facilities and equipment section of the questionnaire, are imputed with the help of a "hot deck" imputation technique that randomly chooses a donor from a group of answering households with similar characteristics.

Income and expenditure data are imputed with the nearest neighbour technique. The imputation is done on one group of variables at a time with the groups being chosen by taking the relationships among variables into account. A group generally corresponds to a section of the questionnaire. For every group, the missing values of the recipients (households that have some missing data for at least one of these variables) are imputed from data of the most similar record among all donors (households that have no missing values for these variables). For each recipient the closest donor is chosen as the one that minimizes a particular distance function. This function is based on matching variables chosen because they are correlated with the variables to be imputed. For example, the total income of a household is chosen as a matching variable for all sections pertaining to expenditures. It must also be ensured that, after receiving the donor values, the recipient household satisfies some consistency rules. In general, the imputation is done at the household level, but in some groups, e.g., income and clothing expenditures, the imputation is done at the person level since the original data is collected at that level.

The bias caused by imputation of partial nonresponse is difficult to evaluate. It depends on the differences between respondents and nonrespondents as well as the ability of the imputation method to produce unbiased estimates. However, the imputation rate gives an indication of the importance of partial nonresponse. Also, the impact of imputed values on the total estimates can be a good indicator

of potential bias in these estimates. These data quality indicators are presented in the following sections[8].

5.1 Impacts and Rates of Imputation by Type of Expenditure and Source of Income

The imputation rate of any variable is defined as the percentage of usable households (or usable persons for appropriate variables) requiring this variable to be imputed. In SHS, the partial nonresponse corresponds mainly to respondents who indicate that they have spent money for a certain category of expenditure but their total amount for the required reference period is unknown. Imputed values then have to be strictly positive. For any infrequent expenditure category or income source where a high proportion of households report a value of 0, the proportion of the estimates that is accounted for by imputed values then becomes a better measure of the imputation effect. This measure, referred to here as the impact of imputation, is defined as the total weighted imputed values divided by the total estimate (sum of weighted values). An algebraic description of the impact of imputation is provided in Appendix 2.

5.1.1 Impact of Imputation on Source of Income and Expenditures by Section

This indicator has the advantage that it can be calculated for aggregate expenditures. The impact on total income is around 0.2% whereas for total expenditures, it is at 0.4%. The largest impacts are on clothing expenditures, miscellaneous expenditures, personal taxes, and personal insurance payments and pension contribution where the impacts are respectively 0.7%, 1.3%, 0.9% and 0.6%. For the rest of the categories, the impact is less than 0.5%.

5.2 Proportion of Households or Persons Requiring Imputation at National and Provincial Level

A preliminary indication of the magnitude of the partial nonresponse is the proportion of households requiring imputation and the number of variables imputed by household. The questionnaire can be divided into two major groups of variables, those collected at the household level and those collected at the individual level such as income and clothing expenditure. For this second type of variable, it is accepted that the respondent provides only the total income or total clothing expenditures if he is unable to provide the breakdowns by source of income or type of expenditure. The level of imputation for the components of income and clothing expenditure is then larger but does not effect the total income, total clothing expenditure or total expenditure.

The percentage of households requiring imputation for household expenditure (excluding clothing expenditures) is presented in the next sub-section. The following sub-section presents the percentage of persons requiring imputation for a clothing expenditure variable and the percentage of persons requiring imputation for an income variable. The results are provided at both the national and provincial or territorial levels. This gives an indication of which provinces or

[8] For operational reasons, these data quality indicators are not available for categorical data such as Household Facilities and Equipment for the 1998 SHS.

territories are more affected by imputation than others, as well as compared to the national level.

5.2.1 Household Expenditure Imputation by Province or Territory

The percentage of usable households that required imputation for an expenditure variable (excluding clothing expenditures) is presented in Table 5.1. The usable households correspond to all sampled households excluding no contact, refusal, incomplete and unbalanced as defined in section 2. The table is broken down by the number of variables (out of 237) imputed for a household.

Table 5.1
Households Requiring Expenditure Imputation by Province and Territory

Province or Territory	Households (%) requiring imputation for **EXPENDITURE VARIABLES** (excluding clothing expenditures)			
	Number of variables imputed (out of 237)			**TOTAL**
	1	**2**	**3 or more**	
Canada	**4.7**	**0.8**	**0.4**	**5.9**
Newfoundland	2.2	0.2	0.1	2.5
P.E.I.	2.9	0.5	0.2	3.6
N.S.	5.1	0.9	0.9	6.9
N.B.	2.8	0.2	0.0	3.0
Quebec	1.7	0.3	0.0	2.0
Ontario	8.0	2.4	1.5	11.9
Manitoba	4.9	0.5	0.6	6.0
Sask.	3.5	0.6	0.1	4.2
Alberta	5.1	0.6	0.2	5.9
B.C.	6.4	0.9	0.4	7.7
Yukon	11.8	0.3	0.0	12.1
N.W.T.	9.5	0.0	0.0	9.5
Nvt.	6.8	2.6	1.1	10.5

Table 5.1 indicates that 6% of households at the national level required some expenditure imputation (when we exclude the clothing section), but with a little bit more than 75% of the total (Canada level) having only one variable imputed. There were very few households at the national level that had more than 1 variable imputed (1.2%). The provincial or territorial rates are generally slightly higher or lower than the 5.9% national rate, but three provinces or territories have more then 10% of households with at least one imputation, they are Ontario (11.9%), Yukon (12.1%) and the Nunavut (10.5%). Quebec and the Atlantic provinces with the exception of Nova Scotia have the lowest imputation rates al being under 4%.

5.2.2 Clothing Expenditure and Income Imputation by Province

Since some respondents provide only totals for clothing expenditure and income variables, a two-stage procedure is used to impute these variables (at the individual level). Individuals who require only imputation of certain components are imputed first, followed by those for which totals are available and require

imputations on all their components (see reference [6] for a more detailed description of this process).

The percentage of usable individuals (persons that are members of usable households) requiring imputation for an income variable are presented by province or territory in Table 5.2. The percentage of persons that had exactly one variable imputed, those that had two or more variables (but not all) imputed and the percentage of persons for which only total income was available and required having all their components imputed are shown. The total percentage of persons requiring some form of income imputation is also provided. The last column of Table 5.2 indicates this total percentage of persons requiring some form of imputation for the clothing expenditure variables.

Table 5.2
Persons requiring Income Imputation and Persons requiring Clothing Expenditure Imputation by Province

Province or Territory	Percentage of persons requiring imputation for INCOME VARIABLES				Percentage of persons requiring imputation for CLOTHING EXPENDITURE VARIABLES (out of 11)
	1 variable imputed	2 or more variables imputed (not all)	All variables imputed (total income known)	**TOTAL** (any form of income imputation)	
Canada	**0.8**	**0.1**	**1.6**	**2.6**	**15.2**
Newfoundland	0.4	0.0	1.4	1.8	8.5
P.E.I.	0.0	0.1	1.5	1.6	21.2
N.S.	0.7	0.1	2.8	3.6	12.4
N.B.	0.1	0.0	3.3	3.5	18.1
Quebec	0.5	0.1	0.8	1.3	25.5
Ontario	1.0	0.1	1.4	2.5	11.0
Manitoba	4.1	0.2	2.0	6.3	13.7
Sask.	0.4	0.0	0.7	1.1	14.6
Alberta	0.0	0.1	1.8	2.0	12.5
B.C.	0.5	0.1	2.2	2.7	15.8
Yukon	0.2	0.0	0.2	0.4	12.0
N.W.T.	0.0	0.4	0.2	0.5	13.9
Nvt.	4.9	0.0	0.1	5.1	15.7

From these results it can be seen that less than 3% of the persons from usable households had some imputation performed on at least one income variable. For almost 60% of the total (Canada level), the respondent provided the total income but all their components had to be imputed. For a very large portion of the remaining persons requiring imputation, only one component of income (one variable) had to be imputed. At the provincial or territorial level, the percentage of persons requiring some imputation on at least one income variable is also low, ranging from a high of just over 6% and 5% respectively for Manitoba and Nunavut to just under or equal to a 0.5% for the Northwest Territories.

From the last column of the table, nearly 15% of persons had some form of imputation done on clothing expenditure variables. The provincial rates are all around the national level except for Newfoundland that is much lower with a rate around 9%, and Prince Edward Island and Quebec, having rates reaching

respectively 21.2% and 25.5%. Nearly all these persons provided total clothing expenditure but needed imputation on components. The individuals requiring imputation on their totals is not presented in this table but represent less than 1% of the imputed individuals in each province or territory. The high level of imputation on clothing components implies that clothing component estimates could be greatly affected by imputation while total estimates will be affected negligibly.

BIBLIOGRAPHY

[1] Methodology of the Canadian Labour Force Survey, Catalogue Number 71-526-XPB.

[2] Labour Force Survey, Quality Report, Surveys 0798 to 1298, Household Survey Methods Division.

[3] Lemaître, G. and Dufour, J. (1987). An Integrated Method for Weighting Persons and Families, *Survey Methodology*, Vol.13, n 2, pp.211-220, Statistics Canada

[4] The Nation: 1996 Census of Population 93F0030XDB96009, Statistics Canada, Electronic Shelf.

[5] The Nation: 1996 Census of Population 93F0029XDB96002, Statistics Canada, Electronic Shelf.

[6] Tremblay, J. and Arsenault, S. (2001). Methodology of the Survey of Household Spending, Household Survey Methods Division, Statistics Canada.

[7] Wolter, K.M. (1985). Introduction to Variance Estimation, Springer-Verlag New-York Inc.

[8] Spending Patterns in Canada (1998), Catalogue Number 62-202-XPE.

APPENDIX A

List of detailed tables available on request

1. Nonresponse Rates (No contact, refusal, Unusable data) by urbanization level and province or territory

2. Nonresponse Rates (No contact, refusal, Unusable data) by community in Territories

3. Impact of Imputed Data on Estimates for all Expenditure Variables

4. Impact of Imputed Data on Estimates for all Income Variables by Province

5. Error Rate with the Suppression Rule in Metropolitan Area

APPENDIX B

Algebraic notation

1. Notation of weights

B_k^{-1} : *Design Weight* : inverse inclusion probability for a given household k

$\boldsymbol{w^{NR}_k}$: *Subweight* : weight adjusted for nonresponse

$\boldsymbol{w^{f}_k}$: *Final Weight:* weight adjusted for nonresponse and controls (regression estimator)

2. Nonresponse adjustment

The Nonresponse adjusted weights for a household k, denoted as $\boldsymbol{w^{NR}_k}$ are

$$w_k^{NR} = p_k^{-1} * \frac{1}{rate_g} \qquad with \qquad rate_g = \frac{\sum_{k=1}^{n_{r,g}} p_k^{-1}}{\sum_{k=1}^{n_{s,g}} p_k^{-1}}$$

where

$n_{r,g}$ is the number of respondents in nonresponse group g,

$n_{s,g}$ is the number of eligible households in the sample in nonresponse group g, and

B_k^{-1} is the design weight assigned to household k.

3. Calculation of the slippage rate

$$rate_c = 100 * \frac{w^{NR}{}_c - t_c}{t_c}$$

Où

$w^{NR}{}_c$ is the subweight adjusted for the nonresponse for the control group c,

t_c is the total for the auxiliary data for the control group c

4. Calculation of the Impact of the Imputation on the Estimate

The impact of the imputation on the estimate is defined as

$$I = \frac{\sum_{k=1}^{n_{s,I}} w_k^f \hat{y}_k}{\sum_{k=1}^{n_{s,R}} w_k^f y_k + \sum_{k=1}^{n_{s,I}} w_k^f \hat{y}_k}$$

where

$n_{s,R}$ is the number of usable households who provided a value for the characteristic y,

$n_{s,I}$ is the number of usable households for which value has to be imputed,

y_k is the value provided by the k^{th} household,

y_k is the imputed value for the k^{th} household, and

w^f_k is the final weight of the k^{th} household.