



Catalogue no. 62F0026MIE — No. 001
ISSN: 1708-8879
ISBN: 0-662-35536-9

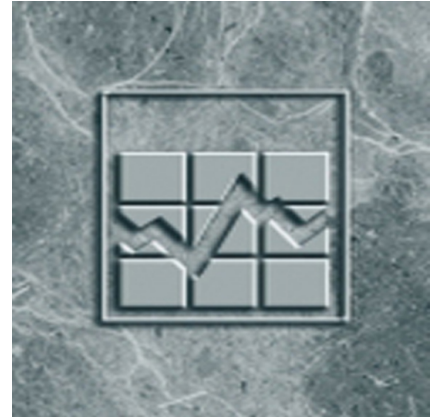
Research Paper

Household expenditures research paper series

Survey of Household Spending 2001 - Data Quality Indicators

2001

Household Survey Methods Division
R.H. Coats Building, Ottawa, K1A 0T6
Telephone: 613 951-7355



Statistics
Canada

Statistique
Canada

Canada

How to obtain more information

Specific 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).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our Web site.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Web site	www.statcan.ca

Ordering and subscription information

This product, Catalogue no. 62F0026MIE2003001, is available on Internet 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 Statistics Canada toll free at 1 800 263-1136.



Statistics Canada
Income Statistics Division

Household expenditures research paper series

Survey of household spending 2001 - Data quality indicators

2001

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2003

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission from Licence Services, Marketing Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

November 2003

Catalogue no. 62F0026MIE2003001

ISSN: 1708-8879

ISBN: 0-662-35536-9

Frequency: Irregular

Ottawa

La version française de cette publication est disponible sur demande (n° 62F0026MIF au catalogue).

Note of appreciation

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

This report was prepared by the members of the Survey of Household Spending unit of the Household Survey Methods Division (HSMD):

Sylvie DeBlois
José Gaudet
Bruno Lapierre
Jessica Law
Christian Nadeau
Sylvain Nadon

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

Table of contents

Highlights	5
Introduction	7
1. Sampling errors	8
1.1 Measures of Sampling Error	8
1.2 Coefficients of Variation	8
Table 1.1	9
Coefficients of Variation (%) by Province, Territory and at the National Level for the Estimation of Average Household Expenditures for Several Summary Level Expenditure Categories and for the Estimation of Average Income	9
Table 1.2	10
Coefficients of Variation (%) at the National, Provincial and Territorial Level for Some Characteristics of Dwelling and Household Equipment.....	10
1.3 Model for Deriving an Approximation of the CV.....	10
1.4 Suppression of Unreliable Data in Estimation Tables	12
2. Nonresponse	12
2.1 Nonresponse Rates and Vacancy Rates.....	12
Table 2.1	13
Nonresponse Rates (%) and Vacancy Rates (%) by Province, Territory and at the National Level	13
2.2 Nonresponse According to Urbanization Level.....	14
Table 2.2	14
Nonresponse Rates (%) and Vacancy Rates (%) by Urbanization Level.....	14
2.3 Nonresponse According to Income Strata	15
Table 2.3	15
Comparison of Nonresponse and Vacancy Rates (%) in High-income and Low-income Strata in Relation to Other Strata.....	15
2.4 Adjustment for Nonresponse	16
3. Coverage errors	16
3.1 Undercoverage and Overcoverage: Slippage Rates	17
Table 3.1	18
National Slippage Rates by Age-Sex Group.....	18
Table 3.2	19
Slippage Rates for Provinces and Territories by Age-Sex Group	19
Table 3.3	20
Slippage Rates for Provinces and Territories by Household Size	20
3.2 Adjustment for Coverage Error at the Population Level and Household Level	20
4. Response errors	21

Table of contents (continued)

5. Processing errors.....	23
5.1 Proportion of Households or Persons Requiring Imputation at National, Provincial and Territorial Level	25
Table 5.1	26
Households Requiring Expenditure Imputation by Province and Territory	26
Table 5.2	27
Persons Requiring Income Imputation and Persons Requiring Clothing Expenditure Imputation by Province and Territory	27
 References	 28
 Appendix A Algebraic notation	 29

Highlights

Sampling Errors

- The coefficients of variation (CVs) of the average estimates of total expenditure per household vary between 1.2% and 1.9% for the provinces, and it is 0.7% at the national level. The CVs are higher in the territories, namely 5.8% in the Yukon, 5.7% in the Northwest Territories and 8.1% in Nunavut.
- The coefficients of variation (CVs) of the average estimates for the different summary level expenditure categories are less than or equal to 2.4% at the national level, and generally lower than 6% at the provincial level. The results for the characteristics of dwelling and household equipment are similar. The CVs at the territorial level are generally higher.

Nonresponse

- The nonresponse rate is 23.8%. It is due to refusals (15.2%), to households that could not be contacted (5.2%), and finally to households with data that were considered unusable (3.4%).
- The final nonresponse rate generally increases with the urbanization level. The nonresponse rate is 19.8% at the rural level, and 30.6% in the urban centres of a million residents or more. This tendency is also observed in the collection nonresponse rates.
- Analysis of final response rates in the strata consisting of high and low-income geographic areas created under the sample design indicates that the nonresponse rate in high-income strata (37.5%) is higher than the rates observed in regular strata (23.5%) and in low-income strata (18.6%). The refusal rate for high-income strata is nearly twice the rate of low-income strata.

Coverage Errors

- There is an undercoverage of 6% for households at the national level. Undercoverage of households is also observed for all provinces and territories, with rates varying from 4.1% to 21.6%. The Northwest Territories and Nunavut have the highest undercoverage of households.
- There is an undercoverage of 7.3% for persons at the national level. Undercoverage of persons is also observed for all provinces, with rates varying from 4.7% to 21.3%. The Northwest Territories and Nunavut have the highest undercoverage of persons.
- The national slippage rates for children (0 to 6 and 7 to 17) are quite different from those of other age groups. Overcoverage or a slight undercoverage occurs with respect to children, while there is always undercoverage among adults. The undercoverage rate for all children combined is 0.7%, while it is 9.2% for adults.

Response Errors

- Response errors include, among others, 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. Total interview time varies according to the household characteristics. For some households the interview can take more than five hours. The average length of interview was one hour and ten minutes.

Processing Errors Related to Imputation

i) Expenditure Variables

- 7.2% of respondents required some expenditure imputation (excluding the clothing section) with the majority of them having only one or two fields imputed out of the 251 expenditure variables.
- About 13% of the individuals required imputation for clothing variables. For the majority of these, the respondents provided the totals and only the components were imputed.

ii) Income Variables

- Less than 4% of individuals required imputation for at least one income variable. For about 80% of these, total income was provided by the respondent and imputation was performed to obtain the breakdown by components.

Introduction

The Survey of Household Spending (SHS) is an annual survey that collects data on household income and expenditure using personal interviews. The 2001 SHS sample is made up of 25,740 dwellings¹ distributed throughout the 10 provinces and the three territories. 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. Following a redesign that took place in 1997, this survey replaces the periodic Family Expenditure Survey and the Household Facilities and Equipment Survey (with modifications to questionnaires and samples).

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 2001 Survey of Household Spending. It covers the usual quality indicators that generally help users interpret data, such as coefficients of variation, nonresponse rates, slippage rates and imputation rates.

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. Response errors and processing errors are dealt with in sections 4 and 5 respectively.

This report focuses on data quality. For a detailed description of the methodology of the survey, see reference [1].

¹ From these 25,740 dwellings, we have to identify and exclude the ineligible dwellings (see section 2.1) to obtain the 22,172 households from which we collect data on household income and expenditure.

1. Sampling errors

Sampling errors exist when inferences about the population are drawn from the survey using 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 [2].

The coefficient of variation (CV) is also a frequently used measure of the reliability of an estimate. It simply 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 $(SE/Y) \times 100$.

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 that, for a characteristic observed, the true value for the population lies within the interval. An interval with a confidence level of 95% corresponds to the estimate obtained from the sample ± 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. The CVs of detailed average household expenditure, as well as the CVs of dwelling characteristics and household facilities and equipment, are available at the national and provincial levels in the publication *User Guide—Survey of Household Spending* (see reference [3]).

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 CVs. For most variables, the imputation rates are low (see section 5) and the provided CVs represent good estimates. To assess the reliability of detailed expenditure with a high imputation rate, the CV and the imputation rate should be considered simultaneously.

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

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

Table 1.1
Coefficients of Variation (%) by Province, Territory and at the National Level 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.	N.L.	P.E.I.	N.S.	N.B.	Que.	Ont.	Man.	Sask.	Alta.	B.C.	Y.T.	N.W.T.	Nvt.
Total expenditure	0.7	1.6	1.9	1.2	1.2	1.2	1.4	1.2	1.4	1.4	1.2	5.7	4.6	8.1
Total current consumption	0.6	1.4	1.7	1.1	1.3	1.1	1.3	1.4	1.4	1.5	1.2	5.0	3.7	6.9
Food	0.5	1.1	1.6	1.1	1.1	0.9	1.1	1.2	1.3	1.2	1.2	3.7	2.8	4.7
Shelter	0.8	1.8	3.0	1.6	2.4	1.5	1.4	2.0	1.9	2.1	1.5	5.8	6.8	12.6
Household operation	0.9	2.3	4.1	1.8	2.2	1.8	1.8	2.2	2.2	2.0	1.8	6.3	6.9	7.8
Furnishings	2.4	3.5	5.5	3.4	4.1	3.7	4.7	3.9	3.6	5.8	4.4	9.1	5.2	10.5
Clothing	1.2	2.7	3.0	2.4	1.9	2.2	2.4	2.5	2.7	2.5	2.5	7.2	5.3	7.5
Transportation	1.4	3.2	4.7	3.2	3.3	3.1	2.8	3.8	3.6	3.1	3.1	8.0	4.6	22.8
Health care	1.3	3.1	6.4	2.6	3.0	2.1	3.5	2.7	2.9	2.2	2.7	10.4	10.6	31.3
Personal care	0.9	2.0	2.5	2.0	1.8	2.0	1.8	1.9	2.0	2.1	2.0	5.1	4.0	8.0
Recreation	1.8	3.3	3.9	3.3	3.4	3.1	3.6	4.1	3.3	4.4	4.3	7.7	6.8	9.6
Reading & printed materials	1.6	3.8	4.8	3.6	3.1	3.1	3.3	3.4	3.2	3.3	3.1	12.6	9.1	9.3
Education	3.1	8.3	10.9	10.8	8.0	5.5	5.6	8.5	7.9	5.7	6.4	19.4	24.9	36.3
Tobacco, alcoholic beverages	1.5	4.3	5.6	3.9	5.1	3.0	3.0	4.3	4.5	3.7	4.1	6.5	5.8	8.3
Games of chance (net)	3.4	10.4	12.9	8.4	7.6	6.8	6.5	10.7	13.6	9.8	8.6	22.4	16.6	17.6
Miscellaneous expenditures	3.0	6.9	8.2	6.4	7.5	6.2	6.2	6.1	5.2	5.1	6.2	18.9	10.4	34.0
Personal income tax	1.7	4.3	4.4	2.6	2.1	2.2	3.5	2.5	2.7	3.6	2.8	9.2	8.3	13.7
Personal insurance	1.5	2.4	3.1	2.5	1.7	2.7	3.1	1.9	2.4	2.6	2.5	12.5	6.1	8.5
Gifts and contributions	3.6	6.9	9.7	6.9	10.0	8.7	6.1	5.9	6.9	7.9	10.0	13.3	17.4	27.4
Income	0.7	1.4	1.7	1.0	0.8	1.3	1.4	0.9	1.1	1.8	1.0	8.5	6.8	11.4

The coefficients of variation (CVs) of the average estimates of total expenditure per household vary between 1.2% and 1.9% for the provinces and it is 0.7% at the national level. The CVs are higher in the territories, namely 5.8% in the Yukon, 5.7% in the Northwest Territories and 8.1% in Nunavut.

For summary level expenditure categories, the CVs at the national level are less than or equal to 2.4%, except for the following categories: *education*, *games of chance*, *miscellaneous expenditures* and *gifts of money and contributions*. These expenditure categories represent respectively 1.6%, 0.5%, 1.5% and 2.2 % of total expenditure (data not presented). Also, with the exception of these categories, the CVs are generally less than or equal to 6% at the provincial level.

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 Dwelling and Household Equipment

CATEGORIES	Can.	N.L.	P.E.I.	N.S.	N.B.	Que.	Ont.	Man.	Sask.	Alta.	B.C.	Y.T.	N.W.T.	Nvt.
Owner	0.9	1.6	3.1	1.7	1.7	1.9	1.8	1.9	1.8	1.8	1.6	8.6	13.9	16.0
Renter	1.6	5.1	7.2	4.5	4.5	2.7	3.3	4.7	4.6	4.5	2.9	15.0	13.5	4.1
Washing machine	0.6	0.9	2.0	1.3	1.1	1.0	1.5	1.7	1.2	1.3	1.4	3.0	4.3	5.4
Clothes dryer	0.7	1.1	2.2	1.4	1.3	1.1	1.5	1.7	1.1	1.4	1.3	3.5	4.2	5.7
Dishwasher	1.1	4.3	5.1	3.1	3.2	2.1	2.5	2.8	2.9	2.2	2.0	11.9	13.1	33.6
Freezer	1.1	1.4	2.5	1.9	2.0	2.5	2.3	1.7	1.7	1.8	2.1	5.8	5.1	9.4
Microwave oven	0.4	0.9	1.2	0.7	0.7	0.9	0.7	0.7	0.9	0.6	0.9	2.9	2.9	6.7
Cellular phone	1.2	3.9	5.5	3.4	4.0	2.9	2.1	2.9	2.9	2.1	2.5	17.2	15.1	24.4
CD player	0.7	1.7	3.3	1.6	1.6	1.6	1.4	1.6	1.5	1.1	1.3	4.3	3.1	3.5
Cable TV	1.0	2.1	5.3	2.3	2.4	2.4	2.0	2.3	2.8	1.8	1.6	10.3	11.0	3.5
Home computer	0.9	2.6	4.3	2.3	2.7	2.2	1.7	2.3	2.6	1.7	1.7	8.3	5.5	14.0
Regular tel. connection to a computer (Modem)	1.9	5.1	7.0	4.6	4.0	3.8	3.4	5.4	5.6	4.9	4.4	12.4	11.3	19.2
Use of internet (home)	1.1	3.2	5.4	2.8	3.4	2.5	2.0	2.6	3.3	2.2	2.2	8.5	5.4	17.7
Owned vehicles (one)	1.5	3.1	4.7	3.2	3.4	2.9	3.2	3.2	3.1	3.2	2.6	8.3	6.3	14.7
Owned vehicles (2 or more)	1.4	4.2	4.6	3.4	3.7	3.3	3.0	3.2	2.6	2.4	2.5	9.1	16.4	37.2

The coefficients of variation for the estimates of dwelling characteristics and household equipment are generally below 5% at the provincial level, with a few exceptions in the category *Regular telephone connection to a computer (Modem)*. The CVs are higher in the territories where there is a smaller proportion of equipment.

The coefficients of variation for the dwelling characteristics and the household equipment at the national level are below or equal to 1.9%.

1.3 Model for Deriving an Approximation of the CV

Estimates for different domains of interest (for example, by income quintile) for the summary level expenditure categories are available in the publication *Spending Patterns in Canada* (see reference [4]). Estimates for different domains of interest for the detailed expenditure are available upon request from the Income Statistics Division. (For more details on tables available upon request from the Income Statistics Division, see references [3] or [4].) 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.

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

It is however 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 below.

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 a household type, an income quintile, an urbanization 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 a CV is desired for a domain that is entirely contained within a single province (for example, a metropolitan area), then it is preferable to use the values at the provincial level, since provincial CVs are published for the 2001 SHS (reference [3]). It should be noted that a CV obtained using this approach is only an approximation of the real value.

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 using another method which will generally provide 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 2001 microdata file. This method of approximation can only be used with the microdata file since it is necessary to have data and weights for each household.

The document on data quality for the 1997 SHS contains the results from the performance evaluation of these two CV approximation methods.

1.4 Suppression of Unreliable Data in Estimation Tables

Since the coefficient of variation is an indicator of the reliability of 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 zero.³

It can be shown that CVs are usually below 33% when the number of households reporting an expenditure is greater than 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%. The document on data quality for the 1997 SHS gives the results from 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 provided information proves to be unusable. When 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 the inability to contact the household or to 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 the 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

Since the units selected in the SHS 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, mobile home moved or dwelling converted to business) and vacant dwellings (unoccupied, seasonal or under construction).

Among eligible dwellings, the proportion of households that did not respond to the survey is evaluated next. This is called the collection nonresponse rate. Included are households that refused to participate in the survey and households where no contact

³ In practice, we use the estimate of the proportion of households reporting an expenditure, which is multiplied by the sample size.

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 questionnaires were at least partially completed but which were rejected during data processing. There are two main causes for rejection. First, when many questions on income or 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 level.

**Table 2.1
Nonresponse Rates (%) and Vacancy Rates (%) by Province, Territory and at the National Level**

Province or territory	Vacancy rate	Collection nonresponse rate			Unusable data rate			Final nonresponse rate (at estimation stage)
		TOTAL	No contact	Refusal	TOTAL	Incomplete	Unbalanced	
Canada	11.5	20.4	5.2	15.2	3.4	1.9	1.5	23.8
Newfoundland and Labrador	18.4	16.1	5.4	10.8	2.1	1.1	1.0	18.3
Prince Edward Island	19.3	15.3	4.6	10.7	0.8	0.4	0.4	16.0
Nova Scotia	13.5	21.6	6.7	14.9	4.3	3.0	1.3	25.9
New Brunswick	15.9	18.4	4.5	14.0	2.8	1.4	1.4	21.2
Quebec	6.9	21.1	6.5	14.6	1.5	0.9	0.6	22.7
Ontario	7.7	28.1	7.3	20.8	7.7	4.8	2.9	35.8
Manitoba	13.8	15.4	3.3	12.1	1.8	0.6	1.2	17.2
Saskatchewan	10.8	17.5	4.1	13.4	2.9	1.2	1.7	20.4
Alberta	5.3	19.3	3.2	16.1	3.5	1.5	2.0	22.8
British Columbia	7.7	22.3	3.8	18.5	3.1	1.8	1.3	25.4
Yukon	31.4	24.1	3.8	20.3	3.2	2.3	0.9	27.2
Northwest Territories	13.3	11.7	6.1	5.6	1.5	0.7	0.7	13.1
Nunavut	12.6	15.0	7.7	7.3	0.4	0.4	0.0	15.4

The final nonresponse rate at the national level is 23.8%. It is due to refusals (15.2%), to households that could not be contacted (5.2%), and finally to households for which the data were unusable (3.4%). For most provinces, refusals are the main cause of nonresponse, followed by the households that could not be contacted, and by the households for which the data were unusable. The only exceptions are Ontario and Alberta where the proportion of unusable data is slightly higher than the proportion of households that we were unable to contact.

The final nonresponse rate varies from one province to another. The Northwest Territories has the lowest nonresponse rate at 13.1%. The lowest refusal rate (5.6%) is also observed in the Northwest Territories. Nonresponse rates in Newfoundland and Labrador, Prince Edward Island, Manitoba and Nunavut are smaller than 20% while rates over 20% are observed for the other provinces and territories. Ontario has the highest nonresponse rate at almost 36%. The highest refusal rate (20.8%) is also observed in Ontario.

The vacancy rates are 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 2001 SHS is 11.5%, which is slightly lower than for the Labour Force Survey (LFS) for the same period. Note that the higher vacancy rate (31.4%) observed in the Yukon is due in part to a different sample allocation for 2001 for this territory and to a lesser extent to modifications made by the LFS to the coverage in the Yukon.

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 2.2.⁴

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

Urbanization category	Vacancy rate	Collection nonresponse rate			Unusable data rate			Total nonresponse rate (at estimation stage)
		TOTAL	No contact	Refusal	TOTAL	Incomplete	Un-balanced	
URBAN								
1,000,000 or more	4.1	26.8	7.8	19.0	3.7	2.1	1.6	30.6
500,000 to 999,999	3.9	18.7	3.3	15.3	2.7	1.0	1.7	21.4
250,000 to 499,999	6.3	26.0	6.9	19.2	5.9	3.8	2.2	32.0
100,000 to 249,999	7.9	22.0	5.0	17.0	3.3	2.1	1.1	25.3
30,000 to 99,999	5.3	19.1	3.7	15.4	3.3	2.1	1.1	22.3
Less than 30,000	12.5	17.1	4.6	12.5	3.0	1.6	1.4	20.1
RURAL	23.4	16.6	4.6	12.0	3.2	1.8	1.4	19.8
TOTAL	11.5	20.4	5.2	15.2	3.4	1.9	1.5	23.8

The final nonresponse rate generally increases with urbanization level. According to Table 2.2, the “500,000 to 999,999” group and, to a lesser extent, the “1,000,000 or more” group, go against this rule.

The collection nonresponse rate also increases with urbanization level. There is a big difference of nearly 10% between the urbanization categories “less than 30,000” and “1,000,000 or more”. Refusals account for more than 60% of the total nonresponse at each level of urbanization. Although the urbanization category “500,000 to 999,999” has

⁴ Tables on nonresponse rates by urbanization level and province are available on request from the Household Survey Methods Division.

the lowest total nonresponse rate among urban regions of 30,000 persons or more, this category has the highest proportion of nonresponse (71.5%) due to refusals.

From an examination of the vacancy rate by urbanization level, it appears that the vacancy rate at the rural level (23.4%) is nearly more than two times higher than for low-population urban areas (12.5%). The latter also shows a rate that is higher than for the 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. 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 nine large cities there are strata consisting of geographic areas where the average household income exceeds \$100,000, and in seven large cities there are 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 580 for high income and 160 for low income strata, or less than 3% of the sample), the comparison of nonresponse rates for 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	Incomplete	Unbalanced	
High-income	2.9	31.5	8.8	22.7	5.9	1.8	4.1	37.5
Regular	11.7	20.1	5.0	15.1	3.3	1.9	1.4	23.5
Low-income	5.1	17.9	6.2	11.7	0.7	0.7	0.0	18.6
TOTAL	11.5	20.4	5.2	15.2	3.4	1.9	1.5	23.8

The final nonresponse rate (37.5%) in high-income strata is much higher than in regular strata. The refusal rate for high-income strata is over 20%, which is nearly twice the rate of low-income strata.

Households in low-income strata have a collection nonresponse rate similar to those for the regular strata. However, regular strata have a final nonresponse rate of 26% higher than that for low-income strata. This is mainly explained by the fact that there are almost no unusable data in the low-income strata.

As for the 1997 to 2000 SHS, the vacancy rate is higher for regular strata than for each of the other two strata. In high-income and low-income strata, the vacancy rates are almost the same.

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. The weighted rates differ from the rates presented in this section since the former takes into account the sampling weight of each household. An algebraic description of the nonresponse adjustment is provided in Appendix A.

The weights adjustment 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. One should note that SHS uses the LFS sampling frame.

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 10 provinces. In the Yukon, the coverage of the population went from 81% in 1999 to 88% in 2001. This increase is due to modifications made by LFS to the coverage of this territory. Thus, some communities and unorganized areas that were previously excluded are now included. 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.

⁵ In terms of households, the coverage of the Yukon, the Northwest Territories and Nunavut represents respectively 88%, 93% and 90% of households.

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), and 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 during data collection, due to improper identification of persons as members of the selected household. These errors also contribute to undercoverage or overcoverage.

A good representation of the target population is essential to the production of realistic expenditure estimates. The number of people per household is also an important characteristic in the estimation of average household expenditures. Therefore, it is necessary that the sample not only adequately represent the individuals in the target population, but also the distribution of households according to their size.

In 1999, a weighting strategy that uses new controls was introduced. This method results in a better correction of the representation of the target population by using a more detailed age-sex grouping than was used previously and for which the coverage varies from one group to the other.

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

⁶ The subweight which is the survey weight adjusted for nonresponse is used (see Appendix A).

**Table 3.1
National Slippage Rates by Age-Sex Group**

National Slippage Rates by Age-Sex Group				
	Age	Sex		Total
		Male	Female	
Canada ¹	0-6 years	0.5	5.2	2.8
	7-17 years	-3.5	-1.6	-2.6
	18-24 years	-6.5	-3.2	-4.9
	25-34 years	-13.4	-10.2	-11.8
	35-54 years	-13.6	-7.9	-10.7
	55-59 years	-7.3	-15.9	-11.7
	60-64 years	-12.4	-4.7	-8.4
	65-69 years	-6.4	-0.8	-3.5
	70 years and +	-9.4	-2.5	-5.4
	Total	-9.2	-5.4	-7.3

¹ Excluding the territories.

For the 2001 SHS, the national undercoverage rate was 7.3%. If we analyze Table 3.1 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 of other age groups. Overcoverage or a slight undercoverage occurs with respect to children, while there is always undercoverage among adults. The undercoverage rate for all children combined is 0.7%, while it is 9.2% for adults (data not presented).

The highest national rates occurred among 55 to 59 years old women, and among 25 to 34 years old and 35 to 54 years old men. Except for the category 55 to 59 years old, the undercoverage rate for women is always smaller than the undercoverage rate for men.

As mentioned previously, the SHS uses the LFS sampling frame. Over the same period, the national LFS undercoverage rate was 9.8% (reference [5]). This is slightly higher than the 8.8% SHS rate for those 15 years or older (data not presented).

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

Slippage Rates by Age-Sex Group														
Sex	Age	Newfoundland and Labrador	Prince Edward Island	Nova Scotia	New Brunswick	Quebec	Ontario	Manitoba	Saskatchewan	Alberta	British Columbia	Yukon	Northwest Territories	Nunavut
Male	0-6	-0.1	-21.1	-2.0	-8.1	12.0	0.2	-14.8	6.4	-8.3	-4.7	7.4	-18.3	-19.4
	7-17	-17.4	5.1	-3.3	-7.4	-10.9	7.3	-6.8	1.6	-8.1	-18.6			
	18-24	-27.1	-31.6	-18.0	-23.0	-1.5	1.9	-13.9	-31.6	-21.9	-9.0	-18.1	-27.4	-20.3
	25-34	-43.3	-13.8	-23.8	-25.9	-7.8	-9.6	-17.2	-12.7	-13.7	-25.6			
	35-54	-14.0	-15.5	-11.9	-15.5	-12.5	-11.9	-11.3	-17.5	-13.7	-19.9			
	55-59	-18.6	-11.3	-15.9	-6.0	-1.3	-8.8	-11.8	-24.3	-1.5	-10.2			
	60-64	-0.3	-6.5	-12.1	-38.1	-7.7	-15.7	-16.4	-28.5	-11.5	-4.8			
	65-69	15.5	-15.5	15.2	4.5	-5.4	-11.7	-10.8	-13.1	2.5	-5.1			
	70 +	9.2	-3.9	-11.6	-4.6	-12.7	-12.7	-0.3	2.0	-1.9	-7.3			
	Total	-16.1	-12.6	-11.2	-14.8	-7.5	-6.3	-11.3	-11.9	-11.4	-15.5			
Female	0-6	-21.7	39.3	-1.6	0.8	14.5	3.8	7.9	8.6	4.9	-4.3	10.3	-15.8	-22.8
	7-17	-9.9	-1.9	0.4	-4.5	-4.2	8.3	-10.4	-2.6	-10.8	-16.2			
	18-24	-30.4	-34.8	-18.5	-16.3	14.7	-5.2	-14.9	-24.6	5.6	-17.7	-10.9	-19.0	-13.3
	25-34	-27.3	-8.1	-15.9	-11.8	-2.1	-11.7	-2.0	0.4	-11.9	-18.4			
	35-54	-11.6	-12.0	-6.3	-5.9	-9.9	-3.0	-7.5	-7.4	-8.7	-17.8			
	55-59	-8.0	4.5	-10.1	-10.5	-2.1	-23.3	-27.6	-28.2	-17.6	-20.7			
	60-64	12.4	-9.9	15.9	2.2	-7.0	-7.3	-12.8	-5.6	-1.0	-1.0			
	65-69	-14.3	-23.1	-9.7	6.9	2.8	-4.4	14.7	-9.9	13.8	-5.0			
	70 +	1.9	6.2	-10.1	-12.1	-8.2	1.6	-2.4	-1.0	3.5	-4.3			
	Total	-13.7	-6.0	-7.2	-7.0	-2.8	-3.2	-6.5	-6.2	-5.3	-14.3			
Total		-14.9	-9.3	-9.1	-10.9	-5.2	-4.7	-8.9	-9.1	-8.4	-14.9	-8.6	-21.3	-18.7

We observe a net undercoverage for all provinces and territories, with the rates varying from 4.7% to 21.3%. The magnitude of the undercoverage is greater for the Northwest Territories and Nunavut. However, a low overall rate of undercoverage is not a guarantee of a better coverage. For example, Prince Edward Island's overall slippage rate (-9.3%) concealed the worst case of overcoverage for a provincial age-sex group (39.3% among 0 to 6 years old females) and the third and fourth worst cases of undercoverage (34.8% and 31.6% among 18 to 24 years old women and men respectively). On the other hand, despite an overall slippage rate of -14.9%, British Columbia has only the thirteenth worst case of undercoverage (25.6% among 25 to 34 years old men). We can also see that the pattern of slippage rates differs substantially for age-sex groups from one province to the next.

Table 3.3
Slippage Rates for Provinces and Territories by Household Size

Province or Territory	Slippage Rate			
	Households	One-person households	Two-person households	Three-person and more households
Canada	-6.0	-4.0	-6.1	-7.3
Newfoundland and Labrador	-10.1	0.8	-0.3	-20.3
Prince Edward Island	-6.3	-4.6	6.2	-16.2
Nova Scotia	-7.4	-2.8	-5.8	-11.4
New Brunswick	-6.7	0.2	1.0	-16.4
Quebec	-5.3	-10.9	0.0	-5.4
Ontario	-4.9	0.0	-12.9	-2.1
Manitoba	-4.4	4.8	-6.0	-9.6
Saskatchewan	-7.0	-4.5	-2.0	-13.1
Alberta	-4.1	-0.7	2.6	-10.9
British Columbia	-11.0	-4.1	-9.5	-16.8
Yukon	-12.2	-32.1	-3.9	-4.7
Northwest Territories	-21.6	-17.9	-5.0	-29.6
Nunavut	-17.7 ⁷

Nationally, the number of households was underestimated by 6%. This slippage rate is comparable to the -7.3% slippage rate for individuals.

Among the provinces and territories, there is a substantial variation in the slippage rate for one-person households. This rate ranges from -32.1% for the Yukon to 4.8% in Manitoba. The same phenomenon occurs with respect to two-person and three-person and more households but with less variation. For two-person households, the slippage rate ranges from -12.9% in Ontario to 6.2% in Prince Edward Island. The slippage rate for households of three or more persons ranges from -29.6% in the Northwest Territories to -2.1% in Ontario. For households of three or more persons, there is always undercoverage. Excepting Quebec, Ontario and the Yukon, the greatest undercoverage is observed for households of three or more persons.

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

To correct the coverage problem illustrated in Tables 3.1 and 3.2 and to reduce the resulting bias, the survey data are adjusted during weighting using demographic estimates for the age groups defined in these tables, for each province. For more information on the adjustment's methodology, see reference [1]. This adjustment greatly reduces 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 individuals included for a given age group in a province.

Furthermore, the effectiveness of the coverage adjustment based on demographic estimates depends mostly on the quality of those estimates and their accuracy in

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

representing the survey's target population. 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 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 representativity 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 weights of the SHS to reflect post-censal 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. As in the case of demographic estimates of population, the success of such an adjustment depends on the quality of the supplemental data on the number of households. As such, note that some changes were made to the methodology used in the production of the supplemental data on the number of households for the Northwest Territories and Nunavut. The impact that these changes could have on the outcome of statistical tests comparing estimates from both 1999 SHS and 2001 SHS surveys has not been studied for these two territories.

In addition to demographic estimates of age-sex groups by province, 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 when a respondent 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 the difference between receipts (income and other amounts received by the household) and disbursements (expenditures plus net change in assets and liabilities) for each household. 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 that respondents estimate the frequency of the purchases and the price generally paid in order to derive expenditures for a 12-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 or after 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 decide to 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. 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 five hours.

It should be stressed that questions were added to the 2001 SHS. Among other things, for 2001 only, extra questions were included in the survey so that data from the SHS could be used in the weighting of the Consumer Price Index. This change may affect historical comparisons for a few variables. For example, questions were added under "Personal care preparations" to collect specific information about hair care products, makeup, fragrances, deodorants and oral hygiene products. As a result of these extra questions, respondents may have given more precise information and the increase in the estimate for "Personal care preparations" may have been at least partly caused by an improvement in respondent recall. The effect of additional questions on estimates is difficult to quantify.

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, it is necessary to conduct quite costly special studies in an attempt to measure them. 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. Before 2001, data entry was done with the help of an automated verification system that grouped the questionnaires into batches and chose some questionnaires from each batch to be entered a second time. Any errors found were to be corrected. If the number of errors in a batch was greater than a certain threshold, then the entire batch was submitted for re-entry. Due to the introduction of a new data capture system (BLAISE), there was no questionnaire batch verification procedure this year. However, some edits were implemented in the new data capture system to ensure consistency of data captured. The results of a preliminary data capture study seem to show that data capture error rates of the new system are similar to the ones of the old system.

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 follow some essential consistency rules. Unusual situations that may justify corrections are also identified. This stage of verification is done in Statistics Canada's regional offices in case it is necessary to recontact respondents if some supplementary information is required to resolve inconsistencies in the answers provided. Specially-trained members of the verification teams solve any problems identified. Thereafter, other verification checks are done at the 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 does not 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 chosen by taking the relationships among the variables into account. A group generally corresponds to a section of the questionnaire. For every group, the missing values of a recipient (a household that has 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 that are 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.

Note that in 2001, the imputation of all expenditure and income data was done using a new system based on slightly different methodology from the one used by the previous system. The new system allows a better use of categorical variables as matching fields when selecting a donor. The new system was tested prior to its implementation and the results it gave were similar to those with the old system. Also, results with the new system are similar to those of previous years.

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 rates indicate the importance of partial nonresponse. They are presented in the following section.⁸

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

5.1 Proportion of Households or Persons Requiring Imputation at National, Provincial and Territorial Level

A preliminary indication of the magnitude of 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 the latter, the respondent may provide only the total income or total clothing expenditures if he/she 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 this does not affect 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. After data imputation by the system, some corrections might have been needed on both imputed and non-imputed variables aiming to ensure data consistency. The actual percentages of records requiring this kind of modification are about 1% at the household level and 2% at the person level. The results are provided at the national, provincial and territorial levels. This gives an indication of which provinces are more affected by imputation.

5.1.1 Household Expenditure Imputation by Province and Territory

The percentage of usable households that required imputation for an expenditure variable (excluding clothing expenditures) is presented in Table 5.1. Usable households correspond to all households living in eligible dwellings, excluding households who could not be contacted, who refused to participate in the survey, or who provided unusable data (see section 2.1). The table is broken down by the number of imputed variables (out of 251) for a household.

Table 5.1
Households Requiring Expenditure Imputation by Province and Territory

Province	Households (%) requiring imputation for EXPENDITURE VARIABLES (excluding clothing expenditures)			
	Number of variables imputed (out of 251)			TOTAL
	1	2	3 or more	
Canada	5.4	1.0	0.8	7.2
Newfoundland and Labrador	9.9	0.8	0.2	11.0
Prince Edward Island	2.9	0.5	0.8	4.1
Nova Scotia	7.0	1.2	1.2	9.4
New Brunswick	3.8	0.6	0.5	4.9
Quebec	2.5	0.6	0.4	3.5
Ontario	7.2	2.0	2.0	11.3
Manitoba	4.3	0.7	0.6	5.6
Saskatchewan	6.9	1.0	0.8	8.7
Alberta	3.6	0.7	0.6	4.9
British Columbia	4.4	1.2	0.8	6.4
Yukon	6.0	0.4	0.4	6.8
Northwest Territories	8.9	2.8	0.3	12.0
Nunavut	9.1	1.0	0.0	10.0

Table 5.1 indicates that 7.2% of households required some expenditure imputation (excluding the clothing section) at the national level, but nearly 75% of them had only one variable imputed. There are very few households that had more than one variable imputed (1.8%). At the provincial level, Quebec (3.5%), Prince Edward Island (4.1%) and New Brunswick (4.9%) have the lowest imputation rates. The highest rates are observed in Newfoundland and Labrador (11%), Ontario (11.3%) and the Northwest Territories (12.0%). The low percentage of persons for which some variables had to be imputed, combined with the low number of variables that had to be imputed when imputation was necessary, suggests that the imputed values should not have a strong impact on the estimates.

5.1.2 Clothing Expenditure and Income Imputation by Province and Territory

Since some respondents provide only totals for clothing expenditure and income variables, a two-step procedure is used to impute these variables (at the individual level). Individuals who only require imputation of certain components are imputed first. Then, they are followed by those for which totals are available but imputation on all components is required (see reference [1] for a more detailed description of this process).

The percentage of usable individuals (persons who are members of usable households) requiring imputation for an income variable are presented by province and territory in Table 5.2. The percentage of persons who had exactly one variable imputed, those who had two or more variables (but not all) imputed and the percentage of persons for which only total income was available (and hence 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, but for the clothing expenditure variables.

Table 5.2**Persons Requiring Income Imputation and Persons Requiring Clothing Expenditure Imputation by Province and Territory**

Province	Percentage of persons requiring imputation for INCOME VARIABLES				Percentage of persons requiring imputation for at least one of the 15 CLOTHING EXPENDITURE VARIABLES
	1 income variable imputed	2 or more income variables imputed (not all)	All income variables imputed (total income known)	TOTAL (any form of income imputation)	
Canada	0.5	0.2	2.7	3.5	12.6
Newfoundland and Labrador	0.2	0.1	1.7	2.0	14.8
Prince Edward Island	0.0	0.1	5.1	5.2	13.3
Nova Scotia	0.5	0.3	3.1	3.9	15.2
New Brunswick	0.2	0.2	5.9	6.4	13.1
Quebec	0.3	0.1	1.4	1.8	9.8
Ontario	1.3	0.5	2.4	4.2	13.5
Manitoba	1.2	0.2	2.3	3.7	13.2
Saskatchewan	0.4	0.2	2.0	2.5	11.9
Alberta	0.2	0.3	3.2	3.6	15.1
British Columbia	0.6	0.3	2.9	3.8	9.8
Yukon	0.3	0.0	1.2	1.5	12.8
Northwest Territories	0.0	0.0	1.8	1.8	8.5
Nunavut	0.1	0.0	0.1	0.2	7.6

These results show that less than 4% of persons from usable households had some imputation performed on at least one income variable. For about 80% of them, the respondent gave the total income but all their components had to be imputed. For many of the remaining persons requiring imputation, only one component of income (one variable) had to be imputed. Provincially, the percentages of persons requiring some imputation on at least one income variable are also low, ranging from a low of 0.2% for Nunavut to a high of 6.4% for New Brunswick.

From the last column of the table, it can be seen that about 13% of persons required imputation for at least one of the clothing expenditure variables. The rates at the provincial and territorial levels range from 7.6% (for Nunavut) to around 15% (for Newfoundland and Labrador, Nova Scotia and Alberta). Almost all these people provided their total expenditure on clothing but required imputation of the components. The higher level of imputation required on clothing expenditure components implies that the estimates for these components could be greatly affected by imputation, while the effect on the estimates for total clothing expenditures will be negligible.

References

- [1] Tremblay, J. and Arsenault, S. 2001. *Methodology of the Survey of Household Spending*. Ottawa. Household Survey Methods Division, Statistics Canada.
- [2] Wolter, K.M. 1985. *Introduction to Variance Estimation*. New York. Springer-Verlag.
- [3] Statistics Canada, Income Statistics Division. 2001. *User Guide—Survey of Household Spending*. Catalogue no. 62F0026MIE2002002. Ottawa.
- [4] Statistics Canada, Income Statistics Division. 2001. *Spending Patterns in Canada*. Catalogue no. 62-202-XPE. Ottawa.
- [5] Statistics Canada, Household Survey Methods Division. 2001. *Labour Force Survey, Operations Report*. Survey 200112. Ottawa.

Appendix A

Algebraic notation

1. Nonresponse Adjustment

The subweight (i.e. the design weight adjusted for nonresponse) for a household k , denoted as w_k^{NR} , is

$$w_k^{NR} = \pi_k^{-1} * \frac{1}{rate_g} \quad \text{with} \quad rate_g = \frac{\sum_{k \in S_{g,r}} \pi_k^{-1}}{\sum_{k \in S_{g,r}} \pi_k^{-1} + \sum_{k \in S_{g,nr}} \pi_k^{-1}}$$

where

$S_{g,r}$ is the set of respondents in nonresponse group g ,

$S_{g,nr}$ is the set of nonrespondents (refusals, no contacts, unusable data) in nonresponse group g , and

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

2. Calculation of the slippage rate

The slippage rate for a control group c , denoted as $rate_c$, is

$$rate_c = 100 * \frac{\left(\sum_{k \in S_{c,r}} w_k^{NR} \right) - t_c}{t_c}$$

where

$S_{c,r}$ is the set of respondents in control group c ,

w_k^{NR} is the subweight of household k , and

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