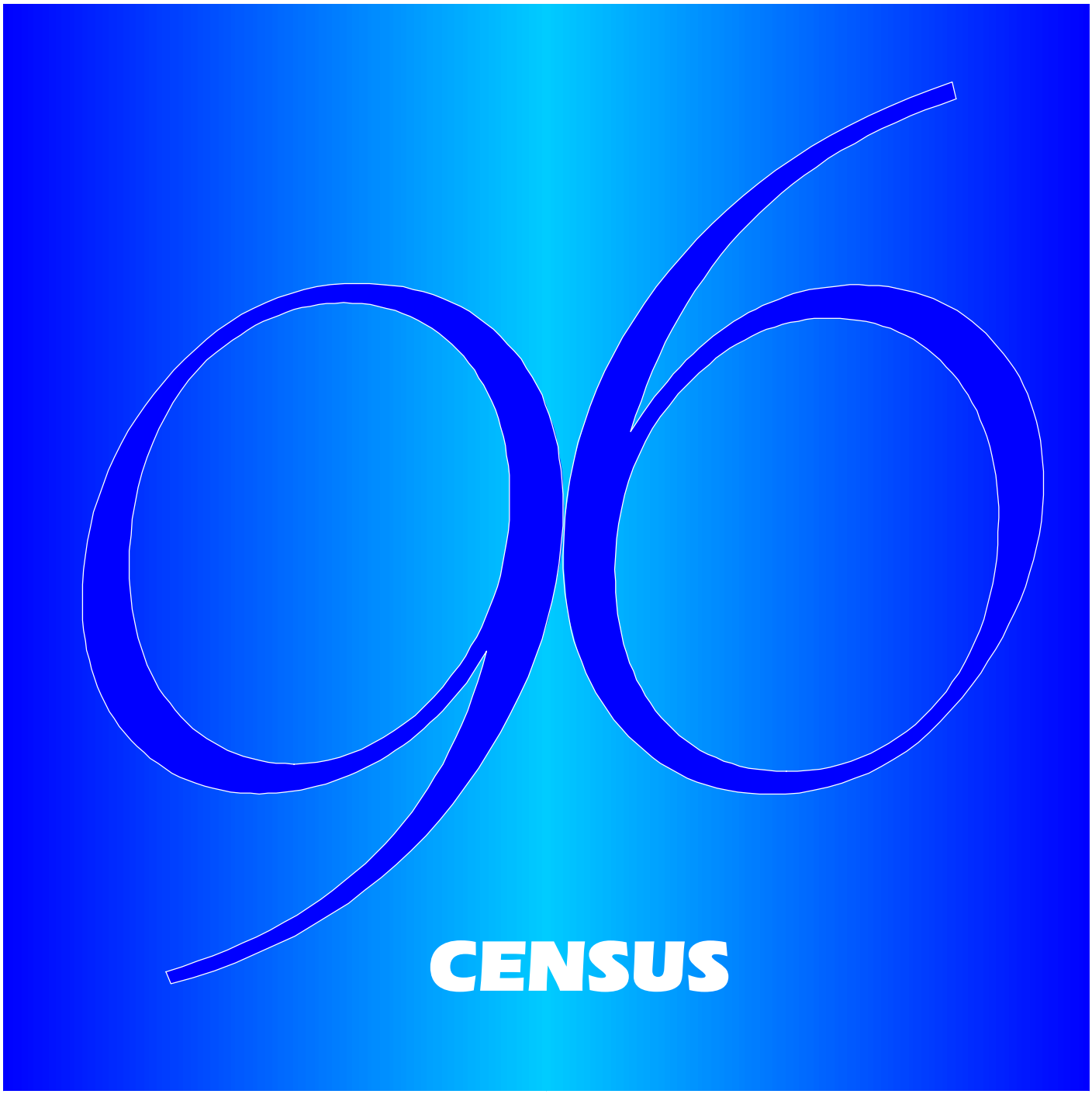
Catalogue No. 92-371-XIE

Sampling and Weighting

1996 Census Technical Reports



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 the Statistics Canada Regional Reference Centre in:

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(780) 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**
Fax order line (Canada and United States)	**1 877 287-4369**

Ordering/Subscription information

Catalogue No. 92-371-XIE is available on Internet free.

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

Sampling and Weighting

1996 Census Technical Reports

Published by authority of the Minister
responsible for Statistics Canada

February 2000

Catalogue No. 92-371-XIE

Ottawa

Note of Appreciation

Canada owes the success of its statistical system to a long-standing cooperation involving 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.

Table of Contents

Page

Page

List of Tables

I. Introduction

The 1996 Census required the participation of the entire population of Canada, i.e. some 29 million people distributed over a territory of 9.2 million square kilometres. An endeavour of this magnitude represented a tremendous challenge. Although there are high quality standards governing the gathering and processing of the data, in spite of efforts aimed at reducing non-response, for example through the use of communications, it is not possible to eliminate all errors. An error does not necessarily imply a mistake as such, as some element of error is bound to result from decisions to control census costs.

Statistics Canada must explain the methods and concepts used to collect and process its data, and provide users with information about the quality of the data produced as well as other data characteristics which might limit their usefulness or interpretation. This report is aimed at informing users about the complexity of the data and any difficulties that could affect their use. It explains the theoretical framework and definitions used to gather the data, and describes unusual circumstances that could affect data quality. Moreover, the report touches upon data capture and edit and imputation, and deals with the chronological comparability of the data.

This report deals with sampling and weighting. It has been prepared by Peter Dick, Karen Switzer, Sylvain Thivierge and Patrick Mason, with the support of staff from two divisions in Statistics Canada: the Social Survey Methods Division and the Census Operations Division.

Users will find additional information on census concepts, variables and geography in the *1996 Census Dictionary* (Catalogue No. 92-351-XPE), and an overview of the complete census process in the *1996 Census Handbook* (Catalogue No. 92-352-XPE).

Sampling is an accepted practice in many aspects of life today. The quality of produce in a market may be judged visually by a sample before a purchase is made; we form opinions about people based on samples of their behaviour; we form impressions about countries or cities based on brief visits to them. These are all examples of sampling in the sense of drawing inferences about the “whole” from information for a “part”.

In a more scientific sense, sampling is used, for example, by accountants in auditing financial statements, in industry for controlling the quality of items coming off a production line, and by the takers of opinion polls and surveys in producing information about a population’s views or characteristics. In general, the motivation to use sampling stems from a desire either to reduce costs or to obtain results faster, or both. In some cases, measurement may destroy the product (e.g., testing the life of light bulbs) and sampling is therefore essential. The disadvantage of sampling is that the results based on a sample may not be as precise as those based on the whole population. However, when the loss in precision (which may be quite small when the sample is large) is tolerable in terms of the uses to which the results are to be put, the use of sampling may be cost effective. Furthermore, the reduction in the scale of a study achieved through using sampling may in fact lead to a reduction in errors from non-sampling sources, thus compensating to some extent for the loss of precision resulting from sampling.

The 1996 Census of Population made use of sampling in a variety of ways. It was used in ensuring that the quality of the Census Representative’s work in collecting questionnaires met certain standards; it was used in the control of the quality of coding responses during office processing; it was used in estimating both the amount of under-coverage and the amount of over-coverage which occurred for different reasons; it was used in evaluating the quality of census data. However, the primary use of sampling in the census was during the field enumeration when all but the basic census data were collected only from a sample of households. This report describes this last use of sampling and evaluates the effect of sampling on the quality of census data.

Chapter II reviews the history of the use of sampling in Canadian censuses and describes the sampling procedures used in the 1996 Census. Chapter III explains the procedures used for weighting up the sample data to the population level and

provides operational and theoretical justifications for these procedures. In Chapter IV, the program of studies designed to evaluate the 1996 Census sampling and weighting procedures is presented, while Chapters V through VIII present the results of these studies. Chapter IX presents some conclusions on the weighting procedures used in 1996 and some suggestions for the 2001 Census.

II. Sampling in Canadian Censuses

In the context of a census of population, sampling refers to the process whereby certain characteristics are collected and processed only for a random sample of the dwellings and persons identified in the complete census enumeration. Tabulations that depend on characteristics collected only on a sample basis are then obtained for the whole population by scaling up the results for the sample to the full population level. Characteristics collected on all dwellings or persons in the census will be referred to as "basic characteristics" while those collected only on a sample basis will be known as "sample characteristics".

A. The History of Sampling in the Canadian Census[1]

Sampling was first used in the Canadian census in 1941. A housing schedule was completed for every tenth dwelling in each census subdistrict. The information from 27 questions on the separate Housing Schedule was integrated with the data in the personal and household section of the Population Schedule for the same dwelling, thus allowing cross-tabulation of sample and basic characteristics. Also in the 1941 Census, sampling was used at the processing stage to obtain early estimates of earnings of wage-earners, of the distribution of the working-age population and of the composition of families in Canada. In this case, a sample of every tenth enumeration area across Canada was selected and all population schedules in these areas were processed in advance.

Again in 1951, the Census of Housing was conducted on a sample basis. This time, every fifth dwelling (those whose identification numbers ended in a 2 or a 7) was selected to complete a housing document containing 24 questions. In the 1961 Census, persons 15 years of age and over in a 20% sample of private households were required to complete a population sample questionnaire containing questions on internal migration, fertility and income. Sampling was not used in the smaller censuses of 1956 and 1966.

The 1971 Census saw several major innovations in the method of census-taking. The primary change was from the traditional canvasser method of enumeration to the use of self-enumeration for the majority of the population. This change was prompted by the results of several studies in Canada and elsewhere (Fellegi, 1964; Hansen et al., 1959) that indicated that the enumerator had the effect of contributing significantly to the variance[2] of census figures in a canvasser census. Thus the use of self-enumeration was expected to reduce the variance of census figures through reducing the effect of the enumerator while, at the same time, giving the respondent more time and privacy in which to answer the census questions – factors which might also be expected to yield more accurate responses.

The second aspect of the 1971 Census that differentiated it from any earlier census was its content. The number of topics covered and the number of questions asked were greater than in any previous Canadian census. Considerations of cost, respondent burden and timeliness versus the level of data quality to be expected using self-enumeration and sampling led to a decision to collect all but certain basic characteristics on a one-third sample basis in the 1971 Census. In all but the more remote areas of Canada, every third private household received the "long questionnaire" which contained all the census questions, while the remaining private households received the "short questionnaire" containing only the basic questions covering name, relationship to head, sex, date of birth, marital status, mother tongue, type of dwelling, tenure, number of rooms, water supply, toilet facilities and certain coverage items. All households in pre-

[1] More detailed information for specific censuses can be found in the Administrative Report, General Review, Summary Guide or Census Handbook of the appropriate census. References to these products can be found at the end of this report.

[2] The "variance" of an estimate is a measure of its precision. Variance is discussed more fully in Chapter VIII.

identified remote enumeration areas and all collective dwellings[3] received the long questionnaire. A more detailed description of the consideration of the use of sampling in the 1971 Census is given in "Sampling in the Census" (Dominion Bureau of Statistics, 1968).

The content of the 1976 Census was considerably less than that of the 1971 Census. Furthermore, the 1976 Census did not include the questions that cause the most difficulty in collection (e.g., income) or that are costly to code (e.g., occupation, industry and place of work). Therefore, the benefits of sampling in terms of cost savings and reduced respondent burden were less clear than for the 1971 Census. Nevertheless, after estimating the potential cost savings to be expected with various sampling fractions, and considering the public relations issues related to a reversion to 100% enumeration after a successful application of sampling in 1971, it was decided to use the same sampling procedure in 1976 as in 1971.

Most of the methodology used in the 1971 and 1976 Censuses was kept for the 1981 Census, except that the sampling rate was reduced from every third occupied private household to every fifth. Studies done at the time showed that the resulting reduction in data quality (measured in terms of variance) would be tolerable, and would not be significant enough to offset the benefits of reduced cost and response burden and would improve timeliness (see Royce, 1983). Twelve questions were asked on a 100% basis and an additional 34 questions were asked in the sample.

The 1986 Census was the first full mid-decade census. It was decided that only a full census could meet the growing need for local labour market data, a need made more pressing by the occurrence of a major recession (1981-82) since the previous census. However, in order to keep development costs as low as possible, a policy of minimum change was adopted. Unless there were compelling reasons not to do so, the 1981 Census questions and data collection and processing procedures were to be retained. Questions on eight subjects from the 1981 Census were not asked in 1986, while three new questions were added.

In 1991, the Census of Population included permanent and, for the first time, non-permanent residents – persons who hold student or employee authorizations, or Minister's permits, or who are refugee claimants. In order to identify non-permanent residents, a new question for the 1991 Census was designed, tested and added. In total, twelve new questions were added while four questions from 1986 were not asked in 1991. Two post-censal surveys were conducted following the collection on the 1991 Census. One of these surveys, the Health and Limitation Survey, collected information of the health and general well-being of Canadians. The second survey, the Aboriginal Peoples Survey, collected information on the Aboriginal population living on and off reserves. In addition, in 1991, there was a significant increase in the automation of data processing as well as in the way the products and services were delivered to the clients.

B. The Sampling Scheme Used in the 1996 Census

A wealth of information was collected from everyone in Canada on Census Day, 1996. Of note, in 1996, Census Day itself was moved from the traditional early June date to May 14, 1996. The bulk of the information was acquired on a sample basis. In all self-enumeration areas, a 1 in 5 sample of private occupied households was selected to receive a long questionnaire (Form 2B), containing all census questions. Basic questions on age, sex, marital status, mother tongue and relationship to the household reference person (Person 1) were asked on 100% basis while additional information on dwelling type, tenure and socio-economic questions were asked on a 20% basis.
All dwellings in those areas enumerated by the canvasser method (generally remote areas or Indian reserves) received the Form 2B. All collective dwellings also received the Form 2B. However, the following persons in collective dwellings were not asked the sample questions:

[3] A collective dwelling is a dwelling of a commercial, institutional or communal nature. Examples include hotels, hospitals, staff residences and work camps.

(a) inmates in correctional and penal institutions or jails;

(b) patients in general hospitals, special care homes and institutions for the elderly, and chronically ill or psychiatric institutions;

(c) children in orphanages and children's homes or young offenders' facilities.

Canadians stationed abroad (generally embassy or Armed Forces personnel) were given a Form 2C, which contained the same questions as the Form 2B except that housing questions were not included. However, questions about the person's usual place of residence in Canada were asked. Information on unoccupied private dwellings was recorded on a Form 2A.

The basic drop-off or delivery procedure required the Census Representative (CR) to pre-plan a route covering all dwellings in his or her enumeration area (EA) and then to visit each dwelling and leave a census questionnaire. The selection of the sample, i.e. the decision as to which type of questionnaire to leave at each occupied dwelling, was facilitated by the Visitation Record (VR), the document in which the CR listed each dwelling in his or her area. This document was printed so that every fifth line was shaded to signify that a Form 2B should be delivered. A random start was implemented by deleting either zero, one, two, three or four lines at the start of the VR according to whether the fifth, fourth, third, second or first dwelling in the EA was to be the first to receive the long questionnaire. Thereafter, the dwelling listed on each shaded line automatically received the long questionnaire. These procedures were spelled out in the CR's Manual and emphasized in his or her training in order to minimize the risk of any deviation from the specified procedure for selecting the sample.

In the 1996 Census, a major test on census operations was conducted in over 400,000 households in Eastern Ontario – including the Ottawa area – with a mail-out/mail-back questionnaire. Basically, all the households in the urban areas were listed on an address register and questionnaires were mailed to these dwellings with the sample households (2B) selected systematically from the Address Register. Respondents were expected to mail back the questionnaires; a small group of highly trained personnel followed up on the non-respondents.

In sampling terminology, the census sample design can be described as a stratified systematic sample of private occupied dwellings using a constant 1 in 5 sampling rate in all strata (EAs). As a sample of persons, it can be regarded as a stratified systematic cluster sample with dwellings as clusters. For a more detailed description of the concepts and terminology of sampling, see Sarndal, Swensson and Wretman, 1992.

C. Processing the Census Sample

Once the CR had obtained the completed questionnaire (Form 2A or 2B) from each dwelling in his or her area, and this work had been approved, the questionnaires were sent to one of seven regional processing sites for manual processing. Complete data for each EA were captured and stored on magnetic tapes. The questionnaires and magnetic tapes were then sent to Head Office Processing in Ottawa. Once there, checks were performed by computer for various inconsistencies in the data which required a manual review of the questionnaire to resolve. After all resulting updates to the data for an EA were completed, the data were reformatted and transferred to Edit and Imputation.
The data were loaded onto Edit and Imputation databases, organized by 2A (100%) and 2B (20%), with five regions for each database. The 2A databases contained the basic demographic characteristics for 100% of the population, while the 2B databases contained the data for the 20% sample questions. The data were processed through a series of customized modules, where all problems of invalid, inconsistent and missing data were resolved. The 2A databases were processed first, and a final 2A Canada Retrieval Database was created.

Once the 100% data were finalized, the data for the 20% sample questions were processed. Non-response 2B records were dropped from the 2B databases. A final 2B Canada Retrieval Database was created which contained both the 100% and 20% data for sampled households and persons only. The weights created using the 100% data (as described in Chapter III) were placed on this database.

III. Estimation from the Census Sample

Any sampling procedure requires an associated estimation procedure for scaling sample data up to the full population level. The choice of an estimation procedure is generally governed by both operational and theoretical constraints. From the operational viewpoint, the procedure must be feasible within the processing system of which it is a part; from the theoretical viewpoint, the procedure should minimize the sampling error of the estimates it produces. In the following two sections, the operational and theoretical considerations relevant to the choice of estimation procedures for the census sample are described.

A. Operational Considerations

Mathematically, an estimation procedure can be described by an algebraic formula that shows how the value of the estimator for the population is calculated as a function of the observed sample values. In small surveys that collect only one or two characteristics, or in cases where the estimation formula is very simple, it might be possible to calculate the sample estimates by applying the given formula to the sample data for each estimate required. However, in a survey or census in which a wide range of characteristics is collected, or in which the estimation formula is at all complex, the procedure of applying a formula separately for each estimate required is not feasible. In the case of a census, for example, every cell of every tabulation based on sample data at every geographic level represents a sample estimate which, under this approach, would require a separate application of the estimation formula. In addition, the calculation of each estimate separately would not necessarily lead to consistency between the various estimates made from the same census sample.

The approach taken in the census therefore (and in many sample surveys) is to split the estimation procedure into two stages: (a) the calculation of weights (known as the weighting procedure) and (b) the summing of weights to produce estimated population counts. Any mathematical complexity is then contained in step (a) which is performed just once, while step (b) is reduced to a simple process of summing weights which takes place at the time a tabulation is retrieved. Note that, since the weight attached to each sample unit is the same for whatever tabulation is being retrieved, consistency between different estimates based on sample data is assured.

B. Theoretical Considerations

For a given sample design and a given estimation procedure, one can, from sampling theory, make a statement about the chances that a certain interval will contain the unknown population value being estimated. The primary criterion in the choice of an estimation procedure is the minimization of the width of such intervals so that these statements about the unknown population values are as precise as possible. The usual measure of precision for comparing estimation procedures is known as the standard error. Provided that certain relatively mild conditions are met, intervals of plus or minus two standard errors from the estimate will contain the population value for approximately 95% of all possible samples.

As well as minimizing the standard error, a second objective in the choice of the estimation procedure for the census sample is to ensure, as far as possible, that sample estimates for basic (i.e. 2A) characteristics are consistent with the corresponding known population values. Fortunately, these two objectives are usually complementary in the sense that the sampling error tends to be reduced by ensuring that the sample estimates for certain basic characteristics are consistent with the corresponding population figures. While this is true in general, however, forcing sample estimates for basic characteristics to be consistent with corresponding population figures for very small subgroups can have a detrimental effect on the standard error of estimates for the sample characteristics themselves.

In the absence of any information about the population being sampled other than that collected for sample units, the estimation procedure would be restricted to weighting the sample units inversely to their probabilities of selection (e.g., if all units had a 1 in 5 chance of selection, then all selected units would receive a weight of 5). In practice, however, one almost always has some supplementary knowledge about the population (e.g., its total size and, possibly, its breakdown by a certain variable – perhaps by province). Such information can be used to improve the estimation formula so as to produce estimates with a greater chance of lying close to the unknown population value. In the case of the census sample, a large amount of very detailed information about the population being sampled is available in the form of the basic 100% data at every geographic level. We can take advantage of this wealth of population information to improve the estimates made from the census sample. However, this information can also be an embarrassment in the sense that **it is impossible to make the sample estimates for basic characteristics consistent with all the population information at every geographic level**. Differences between sample estimates and population values become visible when a cross-tabulation of a sample variable and a basic variable is produced. The tabulation has to be based on sample data, with the result that the marginal totals for the basic variable are sample estimates that can be compared with the corresponding population figures appearing in a different tabulation based on 100% data. They will not necessarily agree exactly.

C. Developing an Estimation Procedure for the Census Sample

Given that a weight has to be assigned to each unit (person, family or household) in the sample, the simplest procedure would be to give each unit a weight of 5 (because a 1 in 5 sample was selected). Such a procedure would be simple and unbiased[4] and, if nothing but the sample data were known, it might be the optimum procedure. However, although we know that the sample will contain almost exactly one fifth of all households (excluding collective households and those in canvasser areas), one cannot be certain that it will contain exactly one fifth of all persons, or one fifth of each type of household, or one fifth of all females aged 25-34, and so on. Therefore, this procedure would not ensure consistency even for the most important subgroups of the population. For large subgroups, these fractions should be very close to one-fifth, but for smaller subgroups they could differ markedly from one-fifth. The next most simple procedure would be to define certain important subgroups (e.g., age-sex groups within the province) and, for each subgroup, to count the number of units in the population in the subgroup (N) and the number in the sample (n) and to assign to each sample unit in the subgroup a weight equal to N/n.

For example, if there were 5,000 males aged 20-24 enumerated in Prince Edward Island, and 1,020 of these fell in the sample households, then a weight of 5,000/1,020 = 4.90 would be assigned to each male aged 20-24 in the sample in Prince Edward Island. This would ensure that whenever sex and age in five-year groups were cross-classified against a sample characteristic for Prince Edward Island, the marginal total for the male 20-24 age-sex group would agree with the population total of 5,000. This type of estimation procedure is known as "ratio estimation". It should be noted in this particular example that a weight of 5 would result in a sample estimate of 5,100 (1,020 x 5). The estimation procedure used prior to the 1991 Census was a generalization of ratio estimation called the raking ratio estimation procedure (RREP). For more details on the RREP, see the *User's Guide to the Quality of 1986 Census Data: Sampling and Weighting* (Statistics Canada, 1990) as well as Brackstone and Rao, 1979.

It was decided to use an alternative estimation procedure called the two-step Generalized Least Squares Estimation Procedure (GLSEP) for the 1991 Census. This was done to achieve a higher level of agreement between population counts and the corresponding estimates at the EA level than was possible with the RREP. The standard errors of the estimates under GLSEP for small geographic areas were also reduced. In addition, the GLSEP allowed a single weight to

[4] "Unbiased" means that the average of the estimates obtained by this procedure, over all possible samples, would equal the true population value.

be determined for each sampled household which was used to produce estimates for both person and household characteristics. By contrast, with the RREP, it was necessary to use different weights to produce estimates for household and person characteristics: this contributed to inconsistencies.

With the GLSEP, the initial weights of approximately 5 were adjusted as little as possible for individual households such that there was perfect agreement between the estimates and the population counts for as many of the basic characteristics as possible that are listed in Appendix B (these will be called constraints). It was required that this perfect agreement be achieved at the weighting area (WA) level. Each WA contained on average seven sampled EAs. More information on WAs is given in Chapter VI, Section A of this report. It can be shown that the GLSEP is a regression estimator. For more details on regression estimators, a good reference is Sarndal, Swensson and Wretman, 1992.

D. The Two-step Generalized Least Squares Estimation Procedure (GLSEP)

The weight calculations are carried out independently in each weighting area (WA). Some of the constraints (both at the EA and WA levels) listed in Appendix B have to be discarded for each WA and hence population/estimate agreement cannot be guaranteed for all constraints. Constraints are initially discarded at the WA level because:

– they apply to less than 20 households (these will be called **small** constraints);
– they are redundant (these will be called **linearly dependent** (LD) constraints);
– they are nearly redundant (these will be called **nearly linearly dependent** (NLD) constraints).

A redundant constraint would be, for example, the total number of females, since constraints are already present when using total number of persons and total number of males. A **linearly dependent constraint** occurs when any two (or more) constraints guarantee that the third will be automatically satisfied. An example of a **nearly redundant constraint** can be seen by considering the constraints **marital status = single** and **household size = 1**. If most, but not all, persons in households of size 1 are single, then the two constraints are almost equal and one constraint can be considered NLD. The LD constraints were discarded to increase the computational efficiency of the weighting algorithm. The small and NLD constraints were discarded because, otherwise, the estimates might become unstable and have large standard errors.

After small, LD and NLD constraints are discarded at the WA level, the calculation of the GLSEP weights takes place in two steps. In the first step, the initial weights, which equal the reciprocal of the EA household sampling fraction, are adjusted individually for each EA. Note that some constraints, when they were not discarded at the WA level, may be discarded because of smallness or linear dependence at the EA level. The remaining constraints that have not been discarded in the EA are sorted by the number of households that they apply to at the EA level. The constraints are then split into two groups, with the even-numbered constraints in one and the odd-numbered constraints in the other. The GLSEP weights are calculated at the EA level for each group of constraints. Sometimes, the estimation procedure will produce very small weights (less than zero) or very large weights (greater than 25) in order to obtain the necessary agreement for certain constraints. These weights, which are called “outlier” weights, are considered undesirable. Consequently, when this occurs, the constraints causing them are identified and discarded, and the weights are recalculated. Finally, the weights for the two groups of constraints are averaged together for each sampled household to produce the first step weights for each EA.

The weights produced in the first step are used as initial weights in the second step, where they are adjusted so that agreement is obtained between sample estimates and population counts at the WA level. All constraints not identified as small, LD or NLD at the WA level are used. If any outlier weights are produced, the constraints causing them are identified and discarded, and the final weights are recalculated. Although the second step distorts somewhat the agreement obtained for estimates at the EA level in the first step, the final EA level estimates are still closer to the population counts than they would have been had the first step not been done. For a more detailed explanation of the calculation of the weights, see Bankier, Rathwell and Majkowski, 1992.

In 1996, two separate estimation runs were completed. The second estimation run was deemed necessary when some unacceptable discrepancies were detected for the variable "common-law status" and for "household size" from the first estimation run. The second run included a new constraint "common-law status (equals yes)" and an adjustment to the initial weight. For the first run, the initial household weight equalled the inverse of the household level sampling fraction. However, in the second run, these initial weights were adjusted so that the estimated number for each household size agreed with the population count at the WA level. Further information can be found in Bankier, Houle and Luc, 1997.

GLSEP weights were calculated only for private households that received the long census questionnaire in a sampled EA (1/5 of the private household*s* were sampled, the other 4/5 were not). Private households that received a short questionnaire in a sampled EA received a weight of 0. All private households in non-sampled EAs received a weight of 1, because 100% of the respondents in such areas provide information for the Form 2B. Collective households also received a weight of 1.

IV. The Sampling and Weighting Evaluation Program

The Sampling and Weighting Evaluation Program was designed to determine the effect of sampling and weighting on the quality of census sample data. To this end, four studies were carried out to measure the quality of the census sample data and estimates and to provide information relevant to the planning of future censuses. These studies were:

(a) an examination of sampling bias;

(b) an evaluation of the weighting procedures;

(c) an evaluation of sample estimate and population count consistency;

(d) a study to evaluate the sampling variance for various 20% sample characteristics.

In the remainder of this chapter, these four studies are briefly described. Chapters V through VIII present the results of these studies.

A. Sampling Bias Study

Bias can be introduced into responses to any survey from a number of sources. The objective of this study was to determine if responses to basic questions on Forms 2B were biased in any way and to identify, if possible, the causes of any observed bias.

B. Evaluation of Weighting Procedures

The objective of this study was to evaluate the performance of the GLSEP. The level of agreement between the sample estimates and population counts for the constraints over all WAs in Canada was examined. The number and type of constraints discarded at the WA level as well as the reason for them being discarded were studied to explain observed inconsistencies. In addition, the distribution of the GLSEP weights as well as differences between 1991 results and 1996 results were studied.

C. Sample Estimate and Population Count Consistency Study

This study examined the level of agreement (consistency) between sample estimates and population counts for a wide variety of basic characteristics. This consistency was studied for various geographic areas other than WAs. Comparisons are also made between the consistency achieved in 1996 and 1991 for these characteristics.

D. Sampling Variance

The “variance” of an estimate is a measure of its precision. Estimates of variance for estimators using simple weights of 5 and assuming simple random sampling are relatively inexpensive to calculate. However, estimates of variance for census estimators taking into account the sample design and estimation techniques used are very expensive to calculate. **Adjustment factors** were calculated for the 1986 Census; they are the ratios of the estimates of the standard errors (the square roots of the variances) for census estimates to the simple estimates of the standard errors. An estimate of the

standard error of a census estimate for any characteristic in any geographic area can then be obtained by multiplying the simple estimate of the standard error by the appropriate adjustment factor. In addition, the study discussed how these estimates of the standard error may not be accurate because of the bias introduced into the process by the sample, the data processing and the estimation procedure.

V. Sampling Bias

Estimates based on a sample survey are subject to sampling errors. One type of sampling error arises from the variability in the population. This variability manifests itself in different samples producing different estimates, none of which will necessarily equal the true population value. The estimates will equal the true population value on average **provided that there is no bias in the sample**. With the presence of a bias, the true population value tends to be overestimated or underestimated. Unfortunately, a bias is often difficult to eliminate completely. In the Census of Population, a bias can be introduced into the responses from a variety of sources. These include coverage errors, non-response bias, response bias (e.g., respondents answering differently on the Form 2B than on the 2A), Census Representative (CR) errors (e.g., not selecting the sample according to specifications), processing errors, and so on.

The Census of Population gives a unique opportunity to examine closely the bias of the basic characteristics. Essentially, there are two estimates of the same basic characteristic: the complete population count and the estimate from the 1 in 5 sample estimate. The purpose of the Sampling Bias Study was to examine the bias in the responses to the basic questions on Forms 2B. The sample estimates were produced by multiplying the sample counts at the EA level by simple weights – equal to the inverse of the EA household sampling fraction (approximately 5) – and then summing to the appropriate geographic level[5].

Initially, the results were summed to the national level. In Table 5.1, the national level results for the population count, the estimate, the difference between the estimate and the population count, the discrepancy and the standard error of the estimate for 32 basic characteristics are displayed. In addition, the Z value – which provides a test of statistical significance – is also shown: this test is discussed in more detail below. For 22 characteristics, the resulting Z value is outside the expected range of -2 to 2. For instance, the Z values for males, ages 20-29, ages 75 and over and males over age 15 are all smaller than -2. This indicates that the sample, compared to the population, is under-represented for these groups at the national level. However, the Z values for females, total population, ages 0-14 and married, are greater than 2, indicating that these characteristics are over-represented in the sample. Also, one-person and six-or-more-person households were under-represented while households of size 2 to 5 were over-represented. Generally the sample contains too many children, middle-aged persons, females and married people while it does not contain enough young adults, elderly persons, males and unmarried persons.

Those differences which were statistically significant at the 5% level in Table 5.1 were identified using the Z value

$$Z^{(0)} = \frac{\hat{X}^{(0)} - X}{\sqrt{V(\hat{X}^{(0)})}} \qquad (1)$$

which is discussed in more detail in Appendix C. The statistic was computed using data with imputations for each of the 281 census divisions (CDs) in Canada and for the set of 32 characteristics where $\hat{X}^{(0)}$ is an estimate based on the 2B sample using simple weights while the known 2A population count X and $V(\hat{X}^{(0)})$ is the sampling variance of the estimator $\hat{X}^{(0)}$. The $Z^{(0)}$ values, for the 281 CDs, should approximately follow a normal distribution with mean 0 and variance 1 if a simple random sample – or a systematic sample – of households was selected unbiasedly from each EA and was not affected by processing.

Table 5.1 presents the results at the national level, but this does not explain whether a few negatively-biased results are offsetting other positively-biased results. To examine this question, a summary of the bias statistics in the 281 CDs are

[5] These simple estimates were used instead of the GLSEP estimates because the GLSEP tends to mask the sampling bias by forcing estimates of basic characteristics to equal population counts.

shown in Table 5.2. This table shows the mean, the standard error and the test result (of assessing if the true mean of the Z values is zero) for the Z values. This hypothesis was rejected for any of the 32 basic characteristics, if |T| was over 1.96 (where T is the statistic of the t-test – see Appendix C for more details). Generally, the same characteristics that were detected as being biased in Table 5.1 – at the national level – are also found to be biased when the entire set of CDs is examined. However, certain changes are also notable. For instance, males at the national level were seen to be under-represented in the sample, but when the CDs are examined this bias is not detected. Since Table 5.2 **assumes all CDs are the same size**, we can conclude that larger (urban) CDs have a negatively-biased estimate of males while smaller (rural) CDs probably have an over-representation of males. Conversely, for households of size 5, a bias was detected nationally, but when the sampled CDs are examined, we can see that they are shown to be over-represented in the sample. The explanation is that, while there are very few CDs with either large over- or under-representations, there is a small, but persistent, over-representation throughout the sample.

Table 5.2 also shows the percentage of CDs with Z values less than -2, the percentage greater than 2 and the percentage greater than 0. If the sample was unbiased, we would expect that approximately 2.5% of the CDs would be either less than -2 and a similar percentage greater than 2, while we would expect 50% of the CDs to have Z values larger than 0. Clearly some characteristics demonstrate that the bias is widespread through almost all CDs. For instance, married persons are over-represented in 82% of the CDs. Furthermore, 13% of the CDs have Z values for married persons of over 2, while only 0.4% of the CDs – meaning 1 CD out of 281 – has a Z value less than -2. The census sample clearly contains too many married persons. On the other hand, only 26% of the CDs have an over-representation – a Z value greater than 0 – of one-person households and, in addition, not even one CD had a Z value greater than 2 for one-person households.

As mentioned in the Introduction, there are many possible explanations for the observed differences between the sample estimates based on simple weights and the population counts. One possible cause is the 0.8% of the total households (i.e. 86,183) which were missed/refusal households in the 1996 Census. This compares with 2.6% of the total (i.e. 253,156) in the 1991 Census. These were households which either completely refused to answer the questions or for which the Census Representative (CR) was unable to get any information, because the members of the household were absent during the census-taking period or had moved on or after Census Day without responding. Sometimes the CR was able to determine the number of persons, but usually all other responses had to be imputed for these households. Of the missed/refusal households in 1996, 32,820 (38% of 86,183) were sampled households, while we would have expected that about 20% would be from the sampled households if non-responses were equally likely for 2A and 2B households. In addition, 3,358 of the sampled households, while not missed/refusal (i.e. they provided some responses to the basic questions), provided no answers to the questions asked on a sample basis. During data processing, these 38,820 + 3,358 = 42,179 sampled households with complete non-response to the sampled questions were removed from the sample (i.e. they were converted from Form 2B to Form 2A households so that they became non-sampled households) and the responses to the basic questions only were imputed. This procedure of converting sampled households to non-sampled households is known as 2A/2B document conversion. It is possible that the missed/refusal households and the households without responses to the sample questions had different characteristics (e.g., they could have been smaller) for known responses than other households. Thus converting Forms 2B to Forms 2A could bias the sample. Also, if the imputation system had a tendency to impute certain characteristics for missed/refusal households more often than for other types of households, this would have caused sample estimate and population count discrepancies as well, since only non-sampled households would be affected.

The geographic variation of the bias was also studied. Table 5.3 shows the resulting Z values for 32 characteristics for Canada , the East region, Quebec, Ontario and the West region. While it is difficult to draw definitive conclusions from this table – since the sample sizes vary to a large extent – some patterns become evident. Quebec has a strikingly different pattern for males – the Quebec sample has males over-represented, while the rest of the country has males under-represented in the sample. However, Ontario is different from the other regions in terms of total population. Quebec also has an over-representation in the sample of single persons and common-law status, while the rest of the country has an under-representation; but this might be more a reflection of the difference in the recording of marital

status in Quebec. Notice that the age groups largely agree across the regions, in terms of the direction of the bias except for some minor differences for some middle-aged persons. The sample for persons over age 15 is fairly consistent in all regions, with males being under-represented and females over-represented (except in Ontario, where females are very slightly under-represented). Finally, the household sizes are also fairly consistent in the pattern of bias. Overall, while there are some differences in the regional bias, there are also a lot of patterns that are shared.

Finally, Table 5.4 briefly highlights some of the results from comparing the 1991 Sampling Bias Study with the 1996 study. When comparing the results at the CD level, it can be seen that only for males and single persons did the direction of the bias change. In 1991, these characteristics were over-represented in the sample, while in 1996 they were under-represented. When the number of CDs with Z values greater than 0 are compared, the difference in the males between the two samples is quite striking. In 1991, 63% of the CDs had males over-represented in the sample, while in 1996, 47% were over-represented. The other major change is in the number of CDs under-representing single persons: in 1991, 56% of the CDs over-represented single persons compared to 47% in 1996. However, in terms of ages, the two samples seem to be remarkably similar in pattern – only persons aged 35-44 were over-represented at a higher percentage of CDs in 1991 than in the 1996 sample.

In summary, there are biases present in the 2B sample and they are similar to the biases found in the 1991 Census sample. It is not clear where the bias comes from and it is not clear what can really be done to reduce it. However, the chosen weighting system – GLSEP – will largely correct the biases for those characteristics selected as constraints. It is an open question as to the resulting bias on the characteristics that are not retained as constraints.

Table 5.1 Summary Statistics at the Canada Level

Characteristic	Count	Estimate[1]	Difference[2]	Disc.[3]	S.E.[4]	Z value[5]
Males	13,717,654	13,694,786	-22,868	-0.17	5,752	-3.98
Females	14,176,680	14,222,665	45,985	0.32	5,552	8.28
Total population	**27,894,334**	**27,917,451**	**23,117**	**0.08**	**8,227**	**2.81**
Age 0-4	1,858,332	1,874,111	15,779	0.85	3,073	5.14
Age 5-9	1,932,023	1,950,728	18,705	0.97	3,120	6.00
Age 10-14	1,939,776	1,957,694	17,918	0.92	3,125	5.73
Age 15-19	1,903,023	1,907,732	4,709	0.25	3,074	1.53
Age 20-24	1,840,654	1,816,301	-24,353	-1.32	3,013	-8.08
Age 25-29	1,971,123	1,953,292	-17,831	-0.90	3,053	-5.84
Age 30-34	2,405,559	2,401,580	-3,979	-0.17	3,317	-1.20
Age 35-39	2,486,060	2,482,136	-3,924	-0.16	3,339	-1.18
Age 40-44	2,268,423	2,273,674	5,251	0.23	3,177	1.65
Age 45-49	2,050,229	2,059,233	9,004	0.44	3,040	2.96
Age 50-54	1,581,484	1,589,751	8,267	0.52	2,707	3.05
Age 55-59	1,271,221	1,269,086	-2,135	-0.17	2,448	-0.87
Age 60-64	1,157,926	1,160,459	2,533	0.22	2,338	1.08
Age 65-74	1,991,721	1,996,303	4,582	0.23	3,068	1.49
Age 75 and over	1,236,780	1,225,372	-11,408	-0.92	2,332	-4.89
Single persons	12,779,218	12,741,878	-37,340	-0.29	7,320	-5.10
Married persons	11,537,475	11,628,813	91,338	0.79	6,076	15.03
Widowed persons	1,303,304	1,291,501	-11,803	-0.91	2,130	-5.54
Divorced persons	1,605,136	1,591,530	-13,606	-0.85	2,612	-5.21
Separated persons	669,201	663,729	-5,472	-0.82	1,675	-3.27
Common-law = yes	1,770,338	1,768,774	-1,564	-0.09	3,568	-0.44
Males $\geq$ 15	10,781,073	10,732,804	-48,269	-0.45	4,449	-10.85
Females $\geq$ 15	11,383,130	11,402,113	18,983	0.17	4,006	4.74
One-person households	2,584,348	2,558,041	-26,307	-1.02	2,524	-10.42
Two-person households	3,385,597	3,397,657	12,060	0.36	3,011	4.00
Three-person households	1,804,304	1,809,076	4,772	0.26	2,435	1.96
Four-person households	1,813,493	1,825,159	11,666	0.64	2,378	4.91
Five-person households	737,751	740,921	3,170	0.43	1,640	1.93
Six-or-more-person households	334,207	327,786	-6,421	-1.92	1,124	-5.71

[1] Based on simple weights
[2] Difference: estimate-count
[3] Disc.: discrepancy (100*[estimate-count]/count)
[4] S.E.: standard error of the simple weight estimate
[5] Z value: (estimate-count)/S.E.

Table 5.2 Summary Statistics for Z Values at the Census Division (CD) Level

Characteristic	Mean	Std[1]	T[2]	% CDs (Z > 2)	% CDs (Z < -2)	% CDs (Z > 0)
Males	-0.07	1.08	-1.10	2.1	3.6	47.3
Females	0.55	1.08	8.50	8.2	0.4	69.7
Total population	0.32	1.06	4.98	6.0	0.7	59.8
Age 0-4	0.33	1.06	5.26	6.4	2.5	60.1
Age 5-9	0.33	1.04	5.29	4.3	0.7	61.2
Age 10-14	0.29	0.99	4.87	3.9	0.7	60.1
Age 15-19	0.10	0.98	1.73	3.6	0.7	54.8
Age 20-24	-0.38	1.04	-6.20	1.1	7.1	37.4
Age 25-29	-0.13	1.07	-2.03	1.8	2.8	48.4
Age 30-34	0.15	1.02	2.38	3.2	2.8	60.1
Age 35-39	0.03	0.98	0.60	2.1	1.8	53.7
Age 40-44	0.13	1.01	2.10	3.6	0.7	53.4
Age 45-49	0.22	1.10	3.29	5.3	2.1	59.1
Age 50-54	0.13	1.03	2.12	4.3	2.1	55.5
Age 55-59	-0.05	0.99	-0.78	1.8	2.1	49.8
Age 60-64	0.05	1.10	0.82	2.5	4.6	53.4
Age 65-74	-0.02	1.03	-0.40	1.8	2.1	50.5
Age 75 and over	-0.43	1.01	-7.11	0.4	7.8	31.0
Single persons	-0.13	1.10	-1.91	1.1	3.9	46.6
Married persons	0.87	1.03	14.15	12.5	0.4	81.9
Widowed persons	-0.41	0.96	-7.06	0.4	4.3	31.7
Divorced persons	-0.23	1.02	-3.78	0.7	6.0	42.7
Separated persons	-0.21	0.94	-3.68	1.4	2.1	39.9
Common-law = yes	-0.02	0.91	-0.33	0.7	1.4	52.0
Males $\geq$ 15	-0.39	1.19	-5.50	1.8	6.8	36.7
Females $\geq$ 15	0.34	1.08	5.22	6.4	1.1	62.3
One-person households	-0.69	1.02	-11.37	0.0	9.6	26.3
Two-person households	0.20	1.04	3.18	5.3	2.1	55.9
Three-person households	0.15	0.98	2.54	3.2	0.7	54.8
Four-person households	0.30	1.12	4.44	3.2	2.5	62.3
Five-person households	0.16	0.96	2.74	1.1	1.4	56.9
Six-or-more-person households	-0.26	1.03	-4.18	1.4	5.3	40.9

[1] Std: standard deviation of the Z values

[2] T : Student statistic for testing that mean = 0

Table 5.3 National and Regional Z Values

Characteristic	Canada	East	Quebec	Ontario	West
Males	-3.98	-0.37	2.48	-5.69	-2.91
Females	8.28	3.87	7.57	1.82	4.29
Total population	2.81	2.34	6.94	-2.75	0.89
Age 0-4	5.14	0.33	4.37	2.10	2.91
Age 5-9	6.00	3.75	4.67	1.42	3.26
Age 10-14	5.73	1.61	2.88	2.88	3.81
Age 15-19	1.53	1.03	2.03	0.24	0.15
Age 20-24	-8.08	-2.46	-2.35	-4.78	-5.97
Age 25-29	-5.84	-0.81	-0.26	-6.31	-2.84
Age 30-34	-1.20	2.15	1.78	-3.51	-0.92
Age 35-39	-1.18	0.36	-0.52	-0.98	-0.77
Age 40-44	1.65	1.30	2.10	-0.79	1.30
Age 45-49	2.96	1.47	2.15	1.22	1.31
Age 50-54	3.05	-0.52	2.29	2.19	1.29
Age 55-59	-0.87	1.41	-0.33	-1.96	0.20
Age 60-64	1.08	0.11	3.02	-1.00	0.26
Age 65-74	1.49	-0.68	-1.02	2.93	0.70
Age 75 and over	-4.89	-3.54	-2.73	-1.46	-2.95
Single persons	-5.10	-0.58	0.87	-6.55	-2.69
Married persons	-15.03	6.10	9.16	7.12	8.03
Widowed persons	-5.54	-2.66	-2.01	-2.98	-3.61
Divorced persons	-5.21	-1.98	-0.98	-3.60	-3.85
Separated persons	-3.27	-1.51	-0.30	-2.28	-2.26
Common-law = yes	-0.44	-0.54	1.09	-1.23	-0.74
Males $\geq$ 15	-10.85	-2.31	-1.27	-9.23	-7.19
Females $\geq$ 15	4.74	3.04	6.27	-0.32	1.75
One-person households	-10.42	-5.02	-8.38	-3.20	-5.01
Two-person households	4.00	1.70	0.68	3.32	2.17
Three-person households	1.96	0.15	2.79	0.55	0.26
Four-person households	4.91	2.76	4.32	1.17	2.29
Five-person households	1.93	-0.02	2.43	-0.21	1.71
Six-or-more-person households	-5.71	-0.74	-0.80	-5.18	-3.12

Table 5.4 Comparison Between 1991 and 1996

Characteristic	1996			1991		
	T-Test	Mean	% CD > 0	T-Test	Mean	% CD > 0
Males	-1.1	-0.1	47	5.1	0.3	63
Females	8.5	0.6	70	12.0	0.8	78
Total population	5.0	0.3	60	11.2	0.7	76
Age 0-4	5.3	0.3	60	5.2	0.3	63
Age 5-9	5.3	0.3	61	7.6	0.5	68
Age 10-14	4.9	0.3	60	8.2	0.5	64
Age 15-19	1.7	0.1	55	3.4	0.2	59
Age 20-24	-6.2	-0.4	37	-3.6	-0.2	43
Age 25-29	-2.0	-0.1	48	-2.5	-0.2	46
Age 30-34	2.4	0.2	60	3.2	0.2	60
Age 35-39	0.6	0.0	54	3.0	0.2	62
Age 40-44	2.1	0.1	53	4.3	0.3	61
Age 45-49	3.3	0.2	59	4.3	0.3	62
Age 50-54	2.1	0.1	56	2.3	0.1	58
Age 55-59	-0.8	-0.1	50	-0.1	0.0	51
Age 60-64	0.8	0.1	53	1.0	0.1	51
Age 65-74	-0.4	0.0	51	-0.2	-0.0	49
Age 75 and over	-7.1	-0.4	31	-6.6	-0.4	35
Single persons	-1.9	-0.1	47	3.2	0.2	56
Married persons	14.2	0.9	82	17.4	1.1	85
Widowed persons	-7.1	-0.4	32	-6.1	-0.4	37
Divorced persons	-3.8	-0.2	43	-5.2	-0.3	38
Separated persons	-3.7	-0.2	40	-5.2	-0.3	39
One-person households	-11.4	-0.7	26	-17.1	-1.0	15
Two-person households	3.2	0.2	56	3.1	0.2	58
Three-person households	2.5	0.2	55	3.6	0.2	59
Four-person households	4.4	0.3	62	7.6	0.5	69
Five-person households	2.7	0.2	57	4.6	0.3	58
Six-or-more-person households	-4.2	-0.3	41	-1.4	-0.1	43

VI. Evaluation of Weighting Procedures

A. Weighting Area (WA) Formation

The first stage of the weighting procedures was the formation of WAs. A WA is the smallest geographic area for which agreement for characteristics of the population between certain sample and population counts can be ensured. Larger WAs allow population/estimate consistency to be achieved for more population characteristics but at the expense of small area population/estimate consistency. Therefore, when considering the WA size (number of households), we must acknowledge the trade-off between the need for agreement between the sample estimate and the population counts for small areas, and the need for this agreement for a wide variety of characteristics. For the 1996 Census, a WA was formed by grouping together enumeration areas (EAs) to adhere to the following conditions:

(a) A WA must respect the boundaries of census divisions (CDs).

(b) A WA should contain a population of between 1,000 and 3,000 households.

(c) A WA should, where possible, respect certain high-level boundaries, giving priority in order of importance to census subdivisions (CSDs), then to census tracts (CTs) and last to federal electoral districts (FEDs).

(d) A WA should be made up of contiguous EAs (i.e. be connected*)* and be as compact as possible.

Since the Generalized Least Squares Estimation Procedure (GLSEP) is performed independently within WAs, agreement between sample estimates and population counts is ensured only for those geographic areas which contain only entire WAs. Agreement is not ensured for geographic areas which are completely contained within a part of one WA or which contain parts of different WAs.

The following table shows the size distribution (in households) of the WAs.

Table 6.1 Dwelling Count Distribution and Contiguity of 1996 Weighting Areas

Dwellings	WA
0- 999	4
1,000-1,499	1,686
1,500-1,999	2,213
2,000-2,499	1,417
2,500-2,999	560
3,000+	61
Total	**5,941**

We observe from the above table that 5,876 (98.9%) of the 1996 WAs are within the desired range of 1,000 to 3,000 households. Also, 5,888 (99.1%) of the WAs were contiguous. Most of the remaining WAs were non-contiguous as a result of either an EA or CSD being non-contiguous.

The following table looks at geographic areas in terms of the number of geographic boundaries that are respected by WAs.

Table 6.2 Number of CSDs, CTs and FEDs that Respect WA Boundaries

Description	1996 Geographic Boundaries		
	CSD	CT	FED
Geographic areas containing only one whole WA or part of only one WA	5,373 (89.8%)	4,396 (73.5%)	0 (0%)
Geographic areas containing more than one whole WA	329 (5.5%)	313 (5.2%)	46 (15.6%)
Geographic areas that cross at least one WA boundary	282 (4.7%)	1,272 (21.3%)	249 (84.4%)
Total	**5,984**	**5,980**	**295**

The table indicates that only 282 (4.7%) of the CSDs do not respect WAs, 1,272 (21.3%) of CTs do not respect WAs and 249 (84.4%) of FEDs do not respect WAs.

For more information about weighting areas and their delineation, see Kruszynski, 1999.

B. Evaluation of the Generalized Least Squares Estimation Procedure

The 1991 weighting system was used in 1996 with very few changes to the software. While this implied little overall change to the system, the parameters – such as the defined constraints – could be modified. The aim was to retain a higher proportion of the constraints than were kept in 1991. In addition, due to budgetary reasons, a number of the 100% characteristics used in 1991 were eliminated.

During the first production run, constraints related to household sizes were often dropped at the WA level, thus permitting a population/estimate difference to be generated for these constraints. In addition, there was an important discrepancy of 2.6% for **common-law = yes** (which was not used as a constraint in the first run). These observations suggested that the weighting system was not completely successful in eliminating biases in the census sample. The implication was that the regression weights could then yield biased estimates for 2B characteristics. For these reasons, it was decided to run the weighting system a second time as described in Chapter III, Section D of this report.

To reduce the discrepancy for **common-law = yes**, this characteristic was added as a constraint in the second run. In addition, to help reduce the household size discrepancies, the initial weights were adjusted to agree with the population counts for the various household sizes at the WA level. In the first run, the initial weights for all the households in the same EA was simply the number of households in the EA divided by the number of sampled households in the EA: these weights are the **initial simple weights**. In the second run, the initial simple weights were adjusted to agree with the estimated number of households – for each of the six (1, 2, 3, 4, 5, 6+) household sizes – at the WA level. Thus, initially, prior to running the regression weighting system, there was agreement at the WA level between the sample estimates and the population counts for the six household sizes. These adjusted weights are called **initial poststratified weights**. The regression weighting system was then applied to these weights. Further details can be found in Bankier, Houle and Luc, 1997.

The evaluation of the weighting system is divided into four parts. Firstly, the distribution of the final weights is examined. Secondly, the discrepancies between the population counts and the final regression estimates are displayed and compared to comparable figures from 1991. Thirdly, the constraints that were discarded in the initial run are briefly described, and then complete details on the constraints discarded in the second run are displayed. Finally, the performance of the weighting system is evaluated at the various geographic levels.

1. Distribution of Weights

Table 6.3 shows the distribution of the weights for the two 1996 production runs and the corresponding 1991 distribution. Note that in order to retain more constraints than in 1991, it was decided to permit weights in the range (0.1) which implied that weights in 1996 are in the range of (0.25) instead of the 1991 range of (1.25).

Table 6.3 Distribution of Household Weights

	1991		1996 – Second		1996 – First	
Weight Range	**Number**	**%**	**Number**	**%**	**Number**	**%**
< 1.00	0	-	13,833	0.7	13,611	0.6
1.00 - 1.99	58,934	3.1	70,195	3.4	68,116	3.3
2.00 - 2.99	170,352	8.9	204,749	9.8	198,052	9.5
3.00 - 3.99	338,334	17.7	383,921	18.4	378,736	18.1
4.00 - 4.99	432,603	22.6	461,281	22.1	465,983	22.3
5.00 - 5.99	369,274	19.3	379,204	18.1	388,137	18.6
6.00 - 6.99	242,000	12.7	245,652	11.8	250,521	12.0
7.00 - 7.99	137,472	7.2	143,887	6.9	144,662	6.9
8.00 - 8.99	74,270	3.9	80,913	3.9	80,486	3.9
9.00 - 9.99	39,671	2.1	44,793	2.1	43,947	2.1
10.00 - 14.99	45,808	2.4	56,569	2.7	53,348	2.6
15.00 - 19.99	3,182	0.2	4,660	0.2	4,099	0.2
20.00 - 24.99	357	0.0	525	0.0	484	0.0
≥ 25.00	0	-	0	-	0	-
Total	**1,912,257**	**100.0**	**2,090,182**	**100.0**	**2,090,182**	**100.0**

From this table, it can be seen that only a small proportion of households had weights less than 1; in fact, fewer than 0.7% of households had weights this small. Overall, there was a higher proportion of small and large weights in 1996 than in 1991. Note that, in 1996, every weight size less than 3 has a larger percentage of the total than the corresponding group in 1991. In addition, every weight size of 9 or more also contains a larger percentage than the corresponding group in 1991. However, for the mid-range weights between 4 and 9, 1991 contains a larger percentage of households in this range. When the percentage of weights contained between 3 and 7 is examined from 1986 (Bankier *et al.*, 1992) and from 1991 to 1996, a steep decline is quite apparent. Whereas, in 1986, when the raking ratio estimation procedure (RREP) was used, 93.9% of the households had their weights in this range; in 1991, it dropped to 72.3% and in 1996, to 71.0% for the first production run and to 70.3% for the second production run.

2. Discrepancy Between Population Counts and Sample Estimates

One of the aims of the weighting procedure is to minimize the discrepancies between population counts and the corresponding sample estimates for the constraints. These discrepancies are the result of sampling variability and bias (see Chapter V). Even after the weighting procedure is completed, however, some discrepancies may remain. Discrepancies are measured by the difference between the sample estimate and the population count, expressed as a percentage of the population count, i.e.

$$\textbf{discrepancy} = \frac{\text{sample estimate} - \text{population count}}{\text{population count}} \text{ X } 100 \qquad (2)$$

The numerator of the above expression (sample estimate – population count) is often referred to as the "**difference**". Note that the population count is for sampled EAs only.

Table 6.4 shows the **differences** and **discrepancies** at the Canada level in 1996 for 32 basic demographic characteristics. This table displays the "**differences**" for three separate sample estimates: the initial weights, the weights from the first run and the weights from the second run. Note that the discrepancy has been rounded to two decimal places. The sample estimates and population counts are based on occupied private dwellings in sampled EAs.

The table clearly shows that the first run of the weighting system reduced the difference for most of the characteristics. For instance, the initial weights produced a discrepancy of 91,338 between the sample estimate and the population count for married persons, while the first run of the weighting system reduced this discrepancy to 10. The discrepancies for the younger age groups have also improved substantially. However, the weighting system did not improve the discrepancy for all characteristics. Indeed, for common-law status, the initial weights produced a difference of -1,404, which the weighting system increased to 46,646. Similar distortions also occurred in household size, especially for household size = 5 and household size = 6. Note that, for household size = 5, the initial weights produced a difference of 3,170, which the weighting system increased to a difference of 57,775. These results were the main factor in running the weighting system a second time.

As discussed above, the second run of the weighting system used common-law status as a constraint. In addition, the initial weights were adjusted to agree with the population counts of household size at the WA level. The results of the differences from the second run are also shown in Table 6.4. The main difficulties resulting from the first run were improved by the second run. For instance, the discrepancy for common-law was reduced from 46,646 to 2,415. Household size = 5 had its difference reduced from 57,775 to 27,879. However, some characteristics were made slightly worse by this second run. For example, persons aged 75 and over had a difference of -2,377 in the first run, but this increased to -9,207 with the second run.

Table 6.4 1996 Estimate/Population Discrepancies at the Canada Level

	Initial Weights	First Run		Second Run	
Constraint	**Difference**	**Difference**	**Discrepancy**	**Difference**	**Discrepancy**
Males	-22,868	-	0.00	15	0.00
Males over 14	-48,269	213	0.00	-276	0.00
Persons over 14	-29,285	3	0.00	3	0.00
Total households	1,060	-	0.00	-	0.00
Total population	**23,117**	**-**	**0.00**	**-**	**0.00**
Age 0-4	15,779	34	0.00	-208	-0.01
Age 5-9	18,705	168	0.01	-258	-0.01
Age 10-14	17,918	-205	-0.01	462	0.02
Age 15-19	4,709	1,890	0.10	1,853	0.10
Age 20-24	-24,353	1,501	0.08	803	0.04
Age 25-29	-17,831	-49	-0.00	105	0.01
Age 30-34	-3,979	248	0.01	361	0.02
Age 35-39	-3,924	493	0.02	320	0.01
Age 40-44	5,251	232	0.01	366	0.02
Age 45-49	9,004	1,743	0.09	971	0.05
Age 50-54	8,267	959	0.06	993	0.06
Age 55-59	-2,135	-201	-0.02	254	0.02
Age 60-64	2,533	-3,380	-0.29	3,847	0.33
Age 65-74 #	4,582	-1,056	-0.05	-662	-0.03
Age 75 and over	-11,408	-2,377	-0.19	-9,207	-0.74
Single persons	-37,340	239	0.00	115	0.00
Married persons	91,338	10	0.00	73	0.00
Widowed persons	-11,803	-1,338	-0.10	-1,387	-0.11
Divorced persons	-13,606	951	0.06	1,209	0.08
Separated persons	-5,472	137	0.02	-10	-0.00
Common-law = yes	-1,404	46,646	2.63	2,415	0.14
One-person households #	-	-	-	-4,750	-0.18
Two-person households	12,060	-344	-0.01	-3,331	-0.05
Three-person households	4,772	3,317	0.06	2,614	0.05
Four-person households	11,666	4,822	0.07	6,776	0.09
Five-person households	3,170	57,775	1.57	27,879	0.76
Six-or-more-person households #	-	-65,570	-1.38	-29,187	-0.52

= Constraint added for second run

3. Discarding Constraints

In the first run, a total of 29 constraints (see Appendix B) were used at the WA level, including five-year age groups, marital status, sex and household size. The constraint "Marital status – Separated" was not used because it was linearly dependent on the other marital status constraints. Similarly, the age constraint 60-64 and the household size 6 were not used because they were linearly dependent on other constraints. As discussed above, in order to retain more constraints, the weights in 1996 were allowed to be in the range of (0.25) instead of the range (1.25) used in 1991.

Table 6.5 below shows how often a constraint was discarded for the second run at the WA level and for what reasons. **Linearly dependent (LD)** implies that the constraint is redundant (six-or-more-person households is an example, since total households is a constraint; hence, one of the subgroups has to be redundant) and, as such, the removal will make no difference to the final estimates. **Small** implies that the constraint applies to less than 20 households, while **nearly linearly dependent (NLD***)* and **outlier** implies the constraint causes the weights to be larger than 25 or less than 0.

This table shows how often a constraint was dropped; thus, out of a total of 5,941 WAs, the constraint age 0-4 was dropped 3,154 times. This implies that the potential for a discrepancy between the population count and the sample estimate using the regression weights exists. While there is very little discrepancy for this characteristic – Table 6.4 shows a discrepancy of -208 persons for ages 0-4 – this is because it was a **linearly dependent** constraint. However, for one-person households, the difference increases since it was being dropped for being **nearly linearly dependent**. Note that for one-person households, the constraint was dropped 4,600 times (4,583 times because of NLD) and, consequently, the discrepancy between the population count and the estimate using regression weights is -4,740. There is a clear link between the number of times a constraint is dropped and the resulting difference – provided the constraint is not dropped because of linearly dependence.

One group of characteristics does not follow this pattern. Note that the characteristic age from 0-4, 5-9 and 10-14 are all discarding with fairly high frequency. However, the constraints "persons over 15" and "total persons" are almost always retained. Hence, the larger grouping of age 0-14 is a redundant constraint which is why, in turn, the three age groups are dropped with such regularity for being linearly dependent.

Table 6.5 Frequency of Discarding WA Level Constraints in 1996

Characteristic	Small	LD	NLD	Outlier	Total
Males	0	0	0	1	1
Females*	-	-	-	-	-
Total population	**0**	**0**	**0**	**0**	**0**
Age 0-4	6	3,071	20	57	3,154
Age 5-9	30	709	77	135	951
Age 10-14	35	2,110	33	61	2,239
Age 15-19	6	514	27	96	643
Age 20-24	1	216	133	119	469
Age 25-29	1	347	108	82	538
Age 30-34	1	29	23	42	95
Age 35-39	1	0	6	31	38
Age 40-44	1	3	13	45	62
Age 45-49	1	4	9	50	64
Age 50-54	2	157	67	83	309
Age 55-59	2	636	213	147	998
Age 60-64	3	1,122	973	128	2,226
Age 65-74	4	3	214	81	302
Age 75 and over	36	2,864	100	60	3,060
Single persons	0	0	1	3	4
Married persons	0	0	0	4	4
Widowed persons	2	0	174	345	521
Divorced persons	1	1	213	252	467
Separated persons*	-	-	-	-	-
Common-law = yes	23	0	1	272	296
One-person households	1	12	4,583	4	4,600
Two-person households	0	0	1,154	12	1,166
Three-person households	2	22	189	47	260
Four-person households	23	145	52	37	257
Five-person households	193	997	865	92	2,147
Six-or-more-person households*	-	-	-	-	-
Males aged $\geq$ 15	0	1	136	3	140
Persons aged $\geq$ 15	0	0	1	0	1
Total households	0	0	0	0	0

* Indicates the characteristic was not used as a constraint because it was redundant.

4. Evaluation at Various Geographic Levels

A study was done which compared the absolute differences between sample estimates and population counts for 31 characteristics in 1996 (29 of which were constraints) and 1986 for various geographic levels. The 31 characteristics that were part of this study of absolute differences are listed in Appendix B. The results of the study are summarized in Table 6.6 below. The table contains the percentage of characteristics that had an "R value" within a certain range for the six geographic levels shown in the table. An R value is a ratio between 1996 and 1986 differences obtained from the following equation:

$$R = 100 * \frac{\sum_{1}^{N_{96}} \left| \hat{X}^{96} - X^{96} \right| / \sum_{1}^{N_{96}} X^{96}}{\sum_{1}^{N_{86}} \left| \hat{X}^{86} - X^{86} \right| / \sum_{1}^{N_{86}} X^{86}} \qquad (3)$$

where X^{96} and X^{86} are respectively the 1996 and 1986 population counts for a characteristic. The sample estimate in 1996 based on GLSEP weights is $\hat{X}^{96}$ while the sample estimate in 1986 based on RREP weights is $\hat{X}^{86}$. R values were calculated for each of the six geographic levels (EA, WA, CSD, CD, Province and Canada). The sum of the absolute values of the population/estimate differences were calculated, where N_{96} equals the number of areas for the particular geographic level in 1996 and N_{86} equals the number of areas for the particular geographic level in 1986. An R value in the range of 95 to 105 means that the 1996 estimation system and 1986 estimation system performed equally well. An R value less than 95 means that the 1996 system performed better than the 1986 system for the characteristic at the particular geographic level, while an R value greater than 105 means that it did worse. The results for 1991, when the comparison was made with 62 characteristics (49 of which were constraints), are also shown.

Table 6.6 Percentage of the Characteristics With R Values Falling in Certain Ranges

	R Value	EA	WA	CSD	CD	Province	Canada
1996 vs 1986	< 95	94	64	94	64	52	49
	95-105	0	15	0	15	21	18
	>105	6	21	6	21	27	33
1991 vs 1986	< 95	87	58	81	47	31	29
	95-105	11	11	8	18	14	10
	>105	2	31	11	35	55	61

The 1996 estimation system was effective at reducing the population/estimate differences at all levels compared to the 1986 estimation system. Note that, in 1991, the estimation system improved the differences at the smaller geographic levels – from EA up to CD – but that, at the province and Canada level, the 1986 estimation system had smaller differences.

VII. Sample Estimate and Population Count Consistency

In order for the GLSEP to work well, some of the constraints had to be discarded within each WA before the weights could be calculated. Consequently, many important characteristics were discarded in a number of WAs. As a result, the level of agreement (consistency) between sample estimates and population counts for these characteristics was reduced. Furthermore, many geographic areas of interest do not always consist of complete WAs. Consequently, in these areas the consistency for all characteristics depends on how close the areas come to consisting of complete WAs.

The consistency study examined the discrepancies between sample estimates and population counts (expressed as percentages of the population counts) for the same basic set of 31 characteristics as the Sampling Bias Study (see Appendix B) for the following geographic areas:

(a) census divisions;

(b) census subdivisions;

(c) census tracts;

(d) enumeration areas.

As in Chapter VI (Subsection B.2), the discrepancies between sample estimates and population counts were calculated as:

$$\textbf{discrepancy} = \frac{\text{sample estimate} - \text{population count}}{\text{population count}} \quad \text{X } 100$$

A. Census Divisions (CDs)

The percentiles in Table 7.1 summarize the level of consistency for all sampled CDs in Canada for a wide variety of basic characteristics with a population count[6] greater than 50. Generally, the discrepancies produced for characteristics with population counts ≤ 50 for most geographic areas were found to be relatively large (either positive or negative). Therefore, it was decided to not include geographic areas where the characteristic count was less than or equal to 50, because a few of these areas could significantly alter the percentiles of discrepancies in the tables in this chapter. This alteration would occur if many of these areas had either relatively large positive discrepancies or relatively large negative discrepancies.

In Table 7.1, for each characteristic, N% of the CDs had discrepancies that were less than the Nth percentile while 100 – N% of the CDs had discrepancies that were greater than the Nth percentile. Thus, the discrepancy was between the 10th and 90th percentiles for 80% of the CDs, and the discrepancy was between the 25th and 75th percentiles for 50% of the CDs, etc. For example, the discrepancy for age 0-4 was between -0.44% and 0.05% for 80% of the CDs. A pattern that is symmetric about 0 implies that the difference between the sample estimates and the population counts is spread in an even fashion. However, a non-symmetric distribution implies that sample counts are not evenly spread out between positive and negative differences.

[6] The population count here refers to that of the characteristic. For example, the level of consistency for age 0-4 is summarized for all CDs in which there were more than 50 people in the age group 0-4. The same definition applies to Tables 7.1, 7.2, 7.3 and 7.4.

All CDs consist of complete WAs. Thus the characteristics which were constraints in 1996 and which were rarely or never discarded in a WA had nearly perfect consistency at the CD level[7]. These characteristics include total number of persons, total number of females, number of single persons, and so on. The level of consistency for most of the remaining characteristics, while not perfect, was still quite good, except for those characteristics which represent only a small percentage of the population in most CDs, such as the number of separated persons or the number of six-or-more-person households. A general relationship does exist between the discrepancies and the population counts for all characteristics, in that the consistency improves as the population count for the CD increases.

Table 7.1 also shows the corresponding percentiles of the 1991 discrepancies for CDs. Tables 7.2, 7.3 and 7.4 which follow also contain the 1991 data for the other geographic levels. When comparing the 1991 and 1996 data, the 1996 discrepancies at the CD level are the same as or significantly smaller than the 1991 discrepancies for all of the characteristics in Table 7.1.

The sizes of the discrepancies at the CD level are quite small compared to the discrepancies at the smaller geographic levels that are studied in the following sections.

B. Census Subdivisions (CSDs)

Table 7.2 summarizes the level of consistency between sample estimates and population counts for all sampled CSDs in Canada with a population count for the characteristic greater than 50. It includes the same characteristics as Table 7.1. CSDs do not always consist uniquely of complete WAs. They are also much smaller on average than CDs. Consequently, the consistency was not as good for CSDs as for CDs. In general, as with CDs, the consistency improved as the population count for the CSD increased, for all characteristics. In comparison to 1991, the range of discrepancies between the 10th and 90th percentiles is smaller in 1996 for almost all of the characteristics. This is also true for the 25th to 75th percentile range.

C. Census Tracts (CTs)

Table 7.3 summarizes the level of consistency for all sampled CTs in Canada. As with Tables 7.1 and 7.2, this table only includes CTs where population counts for the characteristic were greater than 50. CTs have larger populations on average than CSDs. The improvement in the level of consistency at the CT level between 1991 and 1996 is very noticeable. The 25th and 75th percentiles for 1996 are equal to zero for almost all characteristics, contrary to the ranges found in 1991. Also, the ranges between the 10th and 90th percentiles are much smaller in 1996.

New specifications were used to create WAs in 1996. These specifications created fewer CT boundary crossings than in 1991. This change could help explain the improvements noted in 1996.

[7] Even for characteristics with perfect consistency, published tabulations of basic characteristics based on sample data will not agree exactly with tabulations of the same characteristics based on 100% data. This is because those residents of collective dwellings who were not asked the sample questions (see Chapter II, Section B) are included in tabulations based on 100% data, but are excluded from tabulations based on sample data.

D. Enumeration Areas (EAs)

EAs are the components of WAs, and WAs are the lowest level at which sample estimates are forced to agree with population counts for most characteristics. EAs are also the components of higher geographic levels (CDs, CSDs, CTs, etc.) and a number of the WAs are, as Table 7.3 shows, components of these higher levels. Consequently, the consistency at the EA level cannot be expected to be as good as it would be at the higher geographic levels that have been studied. Table 7.4 confirms this as it shows that, for most characteristics studied in sampled EAs with a population count for the characteristic greater than 50, the discrepancies are larger than the discrepancies for the geographic levels studied earlier. This is the case in both 1996 and 1991. In comparison to the 1991 discrepancies at the 10th and 90th percentiles (and at the 25th and 75th percentiles), the 1996 discrepancies are lower for most of the characteristics studied. The values for one-person households, however, are very similar *for* 1991 and 1996.

E. Impact of the Changes to the Weighting Procedure in 1996

There are a number of possible explanations for the improvement in consistency found in 1996 compared to 1991:

(a) Fewer constraints were used in 1996. This resulted in fewer constraints being dropped, and allowed the set of constraints used in a particular WA to be very similar to the sets used in other WAs. The consistency between the estimate and the count for a given characteristic is thus more stable among WAs.

(b) In 1996, the weights calculated by the procedure could be smaller than one. This fact also contributed to a reduction in the number of constraints dropped.

(c) Changes in WA formation reduced the number of higher geographic area boundary crossings. Some larger geographic areas now consist of more complete WAs. This contributes to reducing the discrepancies in these areas.

In conclusion, it appears that, for smaller geographic areas, the changes introduced to the weighting procedure for 1996 generally yielded better estimates than in 1991 in the sense that, for areas of the same geographic level, the discrepancy for a given characteristic is generally closer to 0 than for the same characteristic in 1991.

Table 7.1 Percentiles of Sample Estimates and Population Count Discrepancies (as a Percentage of the Population Count) for CDs - 1996 and 1991 Censuses

Characteristics Studied	1996 Percentiles					1991 Percentiles				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Person Characteristics										
Males	0	0	0	0	0	0	0	0	0	0
Females	0	0	0	0	0	0	0	0	0	0
Total population	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Age 0-4	-0.44	0	0	0	0.05	-2.86	-1.07	0	1.06	2.38
Age 5-9	-0.10	0	0	0	0.21	-2.14	-1.05	0	0.63	2.06
Age 10-14	0	0	0	0	0.04	-1.80	-0.56	0	0.99	2.36
Age 15-19	0	0	0	0	0.43	-1.87	0.51	0.55	1.76	3.15
Age 20-24	-0.89	0	0	0.09	1.17	-3.71	-0.95	0.32	2.35	4.14
Age 25-29	-0.95	0	0	0	1.32	-3.07	-1.40	-0.20	0.20	1.78
Age 30-34	0	0	0	0	0	-1.67	-0.39	0	0.62	2.11
Age 35-39	0	0	0	0	0	-3.14	-0.84	0	0.59	2.26
Age 40-44	0	0	0	0	0	-2.33	-0.69	0	1.02	3.04
Age 45-49	0	0	0	0	0	-2.95	-1.14	0	1.80	4.26
Age 50-54	-0.17	0	0	0	0.21	-4.96	-2.14	0.13	2.02	5.03
Age 55-59	-1.05	0	0	0	0.85	-6.13	-2.33	0	1.60	4.21
Age 60-64	-1.51	-0.01	0	0.92	2.52	-3.69	-1.75	0.07	1.93	4.88
Age 65-74	-0.25	0	0	0	0.06	-2.28	-1.08	0	0.59	2.03
Age 75 and over	-3.70	-1.94	-0.17	0	0.81	-7.87	-3.66	-1.07	0.67	4.65
Single persons	0	0	0	0	0	-0.10	0	0	0	0.12
Married persons	0	0	0	0	0	0	0	0	0.08	0.33
Widowed persons	-1.53	-0.33	0	0.09	1.60	-4.22	-2.26	-0.55	0.57	2.27
Divorced persons	-1.00	0	0	0.34	2.04	-4.47	-1.82	-0.14	1.88	4.73
Separated persons	-7.29	-2.25	0	1.15	4.30	-9.33	-3.96	0.44	4.90	11.12
Common-law = yes [1]	-0.86	0	0	0	1.36	-	-	-	-	-
Household Characteristics										
One-person households	-0.57	-0.36	-0.20	-0.10	0.10	-2.54	-1.44	-0.60	0.03	0.73
Two-person households	-0.10	0	0	0	0.01	-0.51	0	0	0.22	0.87
Three-person households	0	0	0	0	0.10	-2.43	-0.92	0.21	1.23	3.17
Four-person households	0	0	0	0	0.16	-1.66	-0.57	0.04	0.92	2.15
Five-person households	-0.59	0	0.60	1.81	3.09	-3.83	-0.59	1.56	4.19	6.79
Six-or-more-person households	-6.55	-3.39	-0.84	0.99	2.65	-14.50	-8.69	-4.20	0.18	5.21

[1] Common-law = yes was not used as a constraint in 1991. Thus, comparison between 1991 and 1996 for this characteristic does not make sense.

Table 7.2 Percentiles of Sample Estimates and Population Count Discrepancies (as a Percentage of the Population Count) for CSDs - 1996 and 1991 Censuses

Characteristics Studied	1996 Percentiles					1991 Percentiles				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Person Characteristics										
Males	-4.27	-1.22	0	1.30	4.17	-5.47	-1.69	0	1.62	5.05
Females	-4.49	-1.33	0	1.22	4.27	-5.49	-1.83	0	1.83	5.56
Total population	**-2.11**	**0**	**0**	**0**	**1.94**	**-3.36**	**-0.37**	**0**	**0.26**	**3.28**
Age 0-4	-16.40	-4.20	0	3.53	15.47	-19.40	-7.14	0	6.23	18.50
Age 5-9	-13.20	-3.45	0	3.66	13.98	-16.50	-6.11	0	5.73	17.01
Age 10-14	-13.00	-4.19	0	3.91	13.02	-17.00	-6.15	0	7.06	18.01
Age 15-19	-12.90	-3.20	0	3.87	13.64	-16.80	-6.03	0	6.92	19.04
Age 20-24	-14.10	-3.97	0	3.47	14.06	-21.00	-7.71	0	7.94	21.04
Age 25-29	-14.20	-3.27	0	4.03	15.15	-20.30	-7.54	0	5.96	19.29
Age 30-34	-12.70	-3.40	0	3.99	12.94	-17.90	-6.21	0	6.02	17.39
Age 35-39	-13.00	-3.43	0	3.70	13.23	-19.20	-6.66	0	5.86	18.71
Age 40-44	-13.50	-3.70	0	3.49	12.81	-19.20	-6.43	0	7.42	19.92
Age 45-49	-13.00	-3.27	0	3.90	13.81	-19.20	-7.37	0	8.45	22.15
Age 50-54	-14.60	-3.46	0	3.27	13.84	-23.10	-9.06	0	8.62	21.29
Age 55-59	-14.80	-2.95	0	4.00	15.20	-22.30	-8.28	0	8.34	22.23
Age 60-64	-15.70	-3,73	0	4.79	15.54	-22.30	-9.43	0	9.14	22.91
Age 65-74	-14.80	-4.25	0	3.91	13.33	-21.20	-8.49	0	6.72	19.36
Age 75 and over	-18.40	-5.91	0	4.14	15.11	-26.20	-12.00	-1.31	7.74	22.02
Single persons	-5.46	-1.61	0	1.83	5.46	-7.26	-2.34	0	2.33	7.14
Married persons	-5.83	-1.83	0	1.73	5.64	-6.33	-1.99	0	2.31	6.98
Widowed persons	-14.40	-4.13	0	3.55	14.45	-18.90	-8.35	0	6.03	17.18
Divorced persons	-14.90	-2.73	0	3.20	15.10	-19.60	-7.28	0	7.72	19.43
Separated persons	-11.10	-0.82	0	0.88	10.51	-20.50	-7.85	0	8.55	22.51
Common-law = yes	-17.40	-4.84	0	3.96	16.12	-	-	-	-	-
Household Characteristics										
One-person households	-11.10	-3.16	-0.14	2.87	10.48	-11.90	-5.07	-0.54	3.51	10.46
Two-person households	-10.80	-3.29	0	3.15	10.09	-11.10	-3.71	0	4.02	11.56
Three-person households	-10.60	-2.56	0	2.82	11.37	-15.30	-5.54	0	6.18	17.21
Four-person households	-10.20	-2.93	0	2.60	10.11	-14.80	-5.03	0	4.44	13.50
Five-person households	-7.89	-0.91	0	3.69	11.42	-14.00	-4.94	0.40	8.11	19.45
Six-or-more-person households	-9.36	-3.29	-0.28	2.12	5.84	-20.80	-10.40	-3.38	4.13	12.07

Table 7.3 Percentiles of Sample Estimates and Population Count Discrepancies (as a Percentage of the Population Count) for CTs - 1996 and 1991 Censuses

Characteristics Studied	1996 Percentiles					1991 Percentiles				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Person Characteristics										
Males	-0.20	0	0	0	0.26	-0.90	0	0	0	0.87
Females	-0.20	0	0	0	0.25	-0.90	0	0	0	0.79
Total population	**0**	**0**	**0**	**0**	**0**	**-0.20**	**0**	**0**	**0**	**0.24**
Age 0-4	-1.40	0	0	0	1.80	-11.00	-1.20	0	1.03	9.77
Age 5-9	-1.90	0	0	0	1.72	-11.00	-1.60	0	0.94	10.20
Age 10-14	-1.90	0	0	0	1.72	-11.00	-1.30	0	2.23	13.00
Age 15-19	-3.10	0	0	0	3.62	-11.00	-2.70	0	3.93	11.60
Age 20-24	-2.10	0	0	0	2.44	-10.00	-1.50	0	1.90	10.80
Age 25-29	-1.90	0	0	0	1.89	-10.00	-1.60	0	0.28	7.31
Age 30-34	-1.10	0	0	0	1.28	-7.80	-0.60	0	0.29	7.46
Age 35-39	-1.00	0	0	0	1.16	-9.30	-1.30	0	0	7.88
Age 40-44	-1.20	0	0	0	1.55	-10.00	-1.40	0	1.02	9.05
Age 45-49	-1.10	0	0	0	1.64	-13.00	-3.10	0	3.06	12.00
Age 50-54	-1.80	0	0	0	2.48	-15.00	-5.00	0	5.84	16.00
Age 55-59	-3.60	0	0	0	3.62	-16.00	-5.00	0	5.84	17.30
Age 60-64	-7.00	0	0	0	9.89	-17.00	-6.50	0	6.54	18.60
Age 65-74	-2.70	0	0	0	2.34	-14.00	-3.00	0	1.61	12.40
Age 75 and over	-14.00	0	0	0	7.31	-22.00	-8.90	0	8.23	21.00
Single persons	-0.30	0	0	0	0.31	-1.20	0	0	0	1.27
Married persons	-0.40	0	0	0	0.35	-1.40	0	0	0	1.59
Widowed persons	-4.40	0	0	0	3.36	-15.00	-5.30	0	3.87	14.90
Divorced persons	-2.20	0	0	0	2.97	-16.00	-5.20	0	4.08	14.20
Separated persons	-4.60	0	0	0	5.53	-24.00	-8.80	0	8.86	25.50
Common-law = yes	-2.50	0	0	0	3.39	-	-	-	-	-
Household Characteristics										
One-person households	-2.80	-0.80	0	0.29	2.06	-7.60	-3.20	-0.40	1.38	5.39
Two-person households	-1.50	0	0	0	1.11	-3.90	-0.30	0	0.40	4.23
Three-person households	-1.60	0	0	0	1.79	-8.90	-2.30	0	3.32	9.95
Four-person households	-0.90	0	0	0	2.01	-7.60	-0.90	0	2.19	9.90
Five-person households	-3.60	0	0	0.12	8.01	-16.00	-4.80	0	7.47	18.50
Six-or-more-person households	-6.00	-2.30	1.04	3.78	7.23	-24.00	-11.00	-0.50	7.11	16.30

Table 7.4 Percentiles of Sample Estimates and Population Count Discrepancies (as a Percentage of the Population Count) for EAs - 1996 and 1991 Censuses

Characteristics Studied	1996 Percentiles					1991 Percentiles				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Person Characteristics										
Males	-4.97	-2.05	0.02	2.14	4.80	-6.25	-2.59	0.07	2.64	6.06
Females	-4.91	-2.08	0	1.96	4.77	-6.29	-2.66	-0.10	2.55	6.24
Total population	**-1.94**	**0**	**0**	**0**	**1.81**	**-3.56**	**0**	**0**	**0**	**3.35**
Age 0-4	-21.40	-10.30	0	10.60	21.78	-26.40	-12.70	-0.22	12.50	26.49
Age 5-9	-17.30	-8.68	-0.10	8.63	17.87	-24.20	-11.50	0.03	11.23	24.00
Age 10-14	-19.20	-9.33	-0.15	9.63	20.01	-25.40	-12.20	0.07	12.23	25.28
Age 15-19	-16.70	-8.26	0.11	8.94	17.86	-25.40	-12.10	0.37	13.15	27.12
Age 20-24	-18.30	-9.30	-0.12	9.39	19.20	-27.90	-13.60	-0.23	13.51	28.48
Age 25-29	-18.40	-9.44	-0.32	9.42	19.36	-26.10	-12.60	-0.51	11.45	25.23
Age 30-34	-17.10	-8.73	-0.14	8.56	17.72	-24.40	-11.60	-0.23	11.43	24.63
Age 35-39	-16.90	-8.53	-0.30	8.22	17.14	-25.10	-12.40	-0.68	11.28	24.60
Age 40-44	-17.20	-9.02	-0.34	8.52	17.92	-26.30	-12.80	-0.37	11.93	26.53
Age 45-49	-17.30	-9.02	-0.40	8.60	18.39	-27.40	-13.10	0	13.65	29.16
Age 50-54	-19.00	-9.59	-0.54	9.37	19.54	-30.00	-15.20	-0.10	14.85	31.77
Age 55-59	-19.90	-10.60	-0.74	9.96	20.66	-29.60	-15.00	0.29	15.76	31.40
Age 60-64	-20.50	-10.30	0.06	10.82	22.07	-30.30	-15.40	-0.50	16.01	31.98
Age 65-74	-17.30	-9.12	-0.40	8.60	17.73	-26.60	-12.60	-0.63	11.72	27.41
Age 75 and over	-18.20	-8.72	-0.06	8.78	18.10	-27.50	-12.70	-0.32	12.50	27.83
Single persons	-6.74	-2.76	-0.39	2.84	6.22	-8.45	-3.43	0.04	3.42	8.03
Married persons	-7.85	-3.39	0.06	3.41	8.08	-8.78	-3.55	0.07	3.76	8.92
Widowed persons	-18.00	-9.39	0.10	8.69	17.55	-23.50	-11.50	-0.60	10.68	22.99
Divorced persons	-18.90	-9.61	0	9.86	19.77	-29.70	-15.20	-0.93	14.50	29.84
Separated persons	-33.10	-20.80	-0.36	11.74	29.78	-28.90	-11.70	3.21	17.38	32.19
Common-law = yes	-6.00	0	0	0	6.66	-	-	-	-	-
Household Characteristics										
One-person households	-13.10	-6.01	0.01	6.00	12.62	-13.50	-6.67	-0.35	5.69	12.38
Two-person households	-13.80	-6.96	-0.23	6.61	13.67	-14.40	-7.08	0	7.06	15.11
Three-person households	-16.00	-8.40	-0.17	7.88	16.21	-22.30	-11.40	-0.12	11.25	23.48
Four-person households	-13.90	-7.04	-0.19	7.07	14.20	-17.60	-8.41	0	8.49	18.45
Five-person households	-16.40	-8.23	-0.10	8.46	16.68	-22.20	-11.00	1.09	13.92	26.33
Six-or-more-person households	-13.30	-5.98	2.01	10.03	18.95	-22.60	-9.30	2.48	14.81	26.63

VIII. Sampling Variance

A sampling error can be divided into two components: variance and bias. The variance measures the variability of the estimate about its average value in hypothetical repetitions of the survey process, while the bias is defined as the difference between the average value of the estimate in hypothetical repetitions and the true value being estimated. The mean square error (MSE) measures the variability of the estimate about the true value in hypothetical repetitions of the survey process. It can be shown that the MSE equals the variance plus the square of the bias. The MSE measures most accurately how far the estimate is from the true population value on average. If the bias is small relative to the variance, however, the variance is a good approximation of the MSE. There is evidence, however, that the bias accumulates as census estimates for progressively larger geographic areas are produced. Thus, the bias can be insignificant for small geographic areas but become large relative to the variance for large geographic areas. Because of this, the variance can be much smaller than the MSE for large geographic areas. The variance of an estimate can be estimated from the sample, but the bias of an estimate cannot. This means that it is not possible to accurately estimate the MSE from the sample unless the bias is small relative to the variance.

In previous censuses, a study to provide estimates of the sampling variance was carried out. A few results from the 1986 study are provided in Section A (for more information, see the *User's Guide to the Quality of 1986 Census Weighting: Sampling and Weighting*). Because at larger geographic levels the bias is felt to be the dominant term in the MSE (see Chapter V), calculating the sampling variance does not provide an accurate estimate of the MSE for large geographic areas; hence, it was decided not to repeat this study for the 1996 Census. A discussion is given in Section B, however, of what impact the estimation methodology used in the 1996 Census had on the sampling variance compared to the 1986 Census estimation methodology.

A. 1986 Census Sampling Variance Study

Chapter V presented results of the Sampling Bias Study, describing the nature and extent of bias in the census sample prior to weighting. Chapters VI and VII presented results on the sampling bias following the application of the weighting procedure. Even with a perfectly unbiased sampling method, the results would still be subject to variance, simply because the estimates are based only on a sample. The variance may be estimated using the data collected by the sample survey.[8] The 1986 Sampling Variance Study was carried out to estimate the effect of the sampling and estimation procedures on those census figures that are based on sample data.

On the basis of the 2B sample data, thousands of tables are produced by Statistics Canada. Conceptually, the estimated sampling variance, which is a measurement of precision, can be associated with every estimate calculated in these tables. This measurement takes into account both the sample design and the estimation method. In practice, however, it cannot be calculated for every census estimate because of high data processing costs. Sampling variance is thus estimated for only a subset of census estimates. From this, the combined effect of the sample design and the estimation method on the sampling variance can be estimated. Simple estimates of sampling variance, which are inexpensive to calculate, can then be adjusted for this impact to produce estimates of sampling variance for any census estimates.

Table 8.1 gives the non-adjusted (simple) standard errors of census sample estimates. The figures in this table were determined by assuming that 1 in 5 simple random sampling and simple weighting by 5 was used. The standard errors are expressed in Table 8.1 as a function of the size of both the census estimate and the geographic area. For example, for an estimate of 50 persons in a geographic area with a total of 500 persons, the non-adjusted standard error is 15. Standard errors are given in Table 8.1 for only a limited number of values for the estimated total and the total number of persons, households, dwellings or families in the area. The following formula may be used to calculate the non-adjusted standard errors for any estimated total for an area of any size:

[8] Unfortunately, the sampling variance does not provide any indication of the extent of non-sampling error.

$$NASE = \sqrt{\frac{4E\ (N - E)}{N}} \qquad (4)$$

where NASE is the non-adjusted standard error, E is the estimated total and N is the total number of persons, households, dwellings or families in the area. For example, for an estimated total of 750 persons in an area with a total of 9,000 persons, the non-adjusted standard error would be:

$$\sqrt{\frac{4(750)(9{,}000 - 750)}{9{,}000}} = 52$$

The 1986 Sampling Variance Study provides adjustment factors[9] by which the non-adjusted standard errors should be multiplied to adjust for the combined effect of the sample design and the estimation procedure. To calculate these adjustment factors, a sample of 401 WAs (out of a total of 5,941 WAs) was selected. The sample was allocated among the ten provinces[10] in such a way as to obtain good estimates of the sampling variance at the provincial level without greatly sacrificing the quality of the estimates at the national level. For each WA in the sample, estimates of the sampling variances for **raking ratio estimates** were calculated for different categories of all the characteristics given in **Table 9 of the *1986 Census User's Guide***: this table is included as Appendix D. (**IMPORTANT NOTE**: These factors were calculated from the 1986 Census, which used the RREP; using this table assumes the adjustment effects are the same for the 1996 Census, even though a different estimation procedure was used.) The estimates of sampling variance at the provincial and national levels were obtained by weighting up the WA level estimates. The adjustment factors for each category of each characteristic were calculated by dividing the square roots of the WA level estimates by the non-adjusted standard errors. Adjustment factors were calculated at the provincial and national levels for each characteristic by averaging the adjustment factors for all of its categories. For further information on how these adjustment factors were calculated, see Béland, 1990.

To estimate the standard error for a given census sample estimate, the adjustment factor applying to the characteristic was determined from Appendix D. The adjustment factor at the national or provincial level for sample characteristics was generally in the range 0.40 to 1.60. Then this factor was multiplied by the non-adjusted standard error selected in Table 8.1.

The following example illustrates how to calculate the adjusted standard errors. Suppose the estimate of interest is the immigrant population in Ontario. The 1986 estimate for this characteristic was 2,081,200. The 1986 Census count for the population of Ontario was 9,001,170. Using equation (4) – which calculates the non-adjusted standard error – results in an estimate of 2,530. From Appendix D, the provincial level adjustment factor for the characteristic "immigrant" is 1.12. Consequently, the adjusted standard error for this estimate is 2,530 x 1.12 = 2,834.

[9] The squares of the adjustment factors are commonly known as "design effects".

[10] The Yukon Territory and Northwest Territories were grouped with British Columbia.

A second example, however, casts doubt on the accuracy of these adjusted standard errors as estimates of the square root of the MSE. The estimated number of persons in the 1986 Census with "Marital status – Married" who lived in private dwellings in sampled EAs was 11,771,126. The number of persons in the 1986 Census who lived in private dwellings in sampled EAs was 24,369,559. Applying equation (4) generates a non-adjusted standard error of 4,934. From Appendix D, the national level adjustment factor for the characteristic "Married" is 0.25. Consequently, the adjusted standard error for this estimate is 4,934 x 0.25 = 1,233. Because marital status is a basic characteristic, however, it is known that the population count of the number of persons in the 1986 Census with "Marital status – Married" who lived in private dwellings in sampled EAs was 11,778,842. The difference between the estimate and the population count is -7,716. The ratio of this difference to the adjusted standard error is -7,716/1,233 = -6.25. A 95% confidence interval for an estimate would normally be defined as plus or minus two times the adjusted standard error. The fact that the ratio of the difference to the standard error is -6.25 suggests that the adjusted standard error of 1,233 is an underestimate of the square root of the MSE.

B. Sampling Variance and Bias With Generalized Least Squares Estimation

In Bankier, Rathwell and Majkowski, 1992, the coefficients of variation (CVs) of the GLSEP for some sample characteristics were compared to the corresponding CVs of the RREP. In both cases, the 1986 Census data were used. The CV of an estimate is the square root of the estimated variance expressed as a percentage of the estimate. For 79 WAs, the estimated CVs were calculated for estimates of 507 EA level and 642 WA level sample characteristics (all of which applied to at least an estimated 60 households in the population). The WA level and EA level estimates were each classified into small estimates (less than or equal to the median value of the estimates) and large estimates (greater than the median value of the estimates). It was found that the median value for the CVs for large WA estimates was 5% for the GLSEP while it was 6% for the RREP. The median value for the CVs for small WA estimates was 13% for the GLSEP while it was 15% for the RREP. The median value for the CVs for large EA estimates was 10% for the GLSEP while it was 12.5% for the RREP. The median value for the CVs for small EA estimates was 15% for the GLSEP while it was 17.5% for the RREP. Thus, there was some reduction in the CVs for the GLSEP compared to the RREP at both the EA and WA levels. Because the variances at higher geographic levels are just the sum of the variances at the WA level, these reductions in the CVs should also hold at higher geographic levels.

Chapter V indicated that the census sample has small but significant biases. These biases are insignificant compared to the sampling variance at the WA level. For higher geographic levels, however, the bias for a characteristic can accumulate if the bias almost always results in overestimates or underestimates. It appears that the effect of the bias is more significant with the GLSEP than with the RREP. This can be seen from Table 6.5 in Chapter VI where the GLSEP has smaller population/estimate differences than the RREP for smaller geographic areas. However, as the geographic area grows, the improvement is not as large, and in 1991 the RREP was superior to the GLSEP at the provincial level. Besides bias introduced by sampling and processing, Bankier, Rathwell and Majkowski, 1992, show in a Monte Carlo study that the GLSEP estimator is biased, though the relative bias is less than 1% for 50% of the characteristics studied. More serious, however, is the fact that the estimated variance of GLSEP estimators has a median relative bias of -25% at the WA level. Thus, they tend to underestimate the true variance. The RREP estimators may suffer from similar biases, but no study of them has been done.

Table 8.1 Non-adjusted Estimates of Standard Errors of Sample Estimates

Estimated Total in the Area	Estimated Total Number of Persons, Households and Dwellings								
	500	**1,000**	**2,500**	**5,000**	**10,000**	**25,000**	**50,000**	**100,000**	**250,000**
50	15	15	15	15	15	15	15	15	15
100	18	19	20	20	20	20	20	20	20
250	22	25	30	30	30	30	30	30	30
500	0	30	40	40	45	45	45	45	45
1,000		0	50	55	60	60	65	65	65
2,500			0	70	85	95	95	100	100
5,000				0	100	130	130	140	140
10,000					0	150	180	190	200
25,000						0	220	270	300
50,000							0	320	400
100,000								0	490
250,000									0

Estimated Total in the Area	Estimated Total Number of Persons, Households and Dwellings					
	500,000	**1,000,000**	**2,500,000**	**5,000,000**	**10,000,000**	**25,000,000**
50	15	15	15	15	15	15
100	20	20	20	20	20	20
250	30	30	30	30	30	30
500	45	45	45	45	45	45
1,000	65	65	65	65	65	65
2,500	100	100	100	100	100	100
5,000	140	140	140	140	140	140
10,000	200	200	200	200	200	200
25,000	310	310	310	320	320	320
50,000	420	440	440	440	450	450
100,000	570	600	620	630	630	630
250,000	710	870	950	970	990	990
500,000	0	1,000	1,260	1,340	1,380	1,400
1,000,000		0	1,550	1,790	1,900	1,960
2,500,000			0	2,240	2,740	3,000
5,000,000				0	3,160	4,000
10,000,000					0	4,900

IX. Conclusion

Sampling is now an accepted and integral part of census-taking. Its use can lead to substantial reductions in costs and respondent burden associated with a census, or alternatively, can allow the scope of a census to be broadened at the same cost. The price paid for these advantages is the introduction of sampling error to census figures that are based on the sample. The effect of sampling is most important for small census figures, whether they are counts for rare categories at the national or provincial level or counts for categories in small geographic areas. It should be noted that response errors and processing errors also contribute to the overall error of census figures, and that it is the same small census figures that are particularly susceptible to the effects of these non-sampling errors. Therefore, even with a 100% census, many small figures would be of limited reliability. As a general rule of thumb for the 1996 Census, figures of size 50 or less that are based on sample data are of very low reliability, while figures up to size 500 tend to have standard errors in excess of 10% of their size.

For many of the characteristics, a certain amount of bias was detected in the sample. A small portion of the bias was found to have been introduced during data processing and edit and imputation. The remaining bias must have been due to one or more factors such as non-response bias, response bias, or the selection of a biased sample by the CRs. The procedures for weighting the sample data up to the population level were carried out successfully, and generally achieved the levels of sample estimate and population count consistency anticipated. While the consistency that was achieved at the provincial and Canada levels was better than in 1991, it is still lower than what might otherwise be expected given the improved consistency for smaller geographic levels that has been achieved. This is probably the result of the accumulation of small biases in the sample summed over many areas.

The census estimation methodology will be reassessed for the 2001 Census to see if it is possible to improve sample estimate and population count consistency at the provincial and Canada levels while maintaining good consistency at the EA level. Doing this should also allow more reliable estimates of the mean square error of the census estimates to be produced.

APPENDICES

Appendix A – Glossary of Terms

The definitions of census terms, variables and concepts are presented here as they appear in the *1996 Census Dictionary* (Catalogue No. 92-351-XPE). Users should refer to the *1996 Census Dictionary* for full definitions and additional remarks related to any concepts, such as information on direct and derived variables and their respective universe.

Census division (CD): Refers to the general term applied to areas established by provincial law which are intermediate geographic areas between the municipality (census subdivision) and the province level. Census divisions represent counties, regional districts, regional municipalities and other types of provincially legislated areas.

Census subdivision (CSD): Refers to the general term applying to municipalities (as determined by provincial legislation) or their equivalent (for example, Indian reserves, Indian settlements and unorganized territories).

Census tract (CT): Small geographic units representing urban or rural neighbourhood-like communities created in census metropolitan areas and census agglomerations (with an urban core population of 50,000 or more at the previous census).

Enumeration area (EA): Refers to the geographic area canvassed by one census representative. It is the smallest standard geographic area for which census data are reported. All the territory of Canada is covered by EAs.

Household: Refers to a person or a group of persons (other than foreign residents), who occupy the same dwelling and do not have a usual place of residence elsewhere in Canada. It may consist of a family group (census family) with or without other non-family persons, of two or more families sharing a dwelling, or a group of unrelated persons, or of one person living alone.

Marital Status: Refers to the conjugal status of a person: Married (including common-law); Separated; Divorced; Widowed; Never married (single).

Occupied private dwelling: Refers to a private dwelling in which a person or a group of persons are permanently residing. Also included are private dwellings whose usual residents are temporarily absent on Census Day.

Private dwelling: Refers to a separate set of living quarters with a private entrance either from outside or from a common hall, lobby, vestibule or stairway inside the building. The entrance to the dwelling must be one than can be used without passing through the living quarters of someone else.

Private household: Refers to a group of persons (other than foreign residents) who occupy a private dwelling and do not have a usual place of residence elsewhere in Canada.

Appendix B – WA and EA Level Constraints Applied to the 1996 Census Weights

(**Note**: The 1996 Census second run additional constraints are flagged with an "#".)

Person WA Level Constraints

– Total persons
– Total persons aged ≥ 15

– Males
– Males aged ≥ 15

– Persons aged 0 to 4
– Persons aged 5 to 9
– Persons aged 10 to 14
– Persons aged 15 to 19
– Persons aged 20 to 24
– Persons aged 25 to 29
– Persons aged 30 to 34
– Persons aged 35 to 39
– Persons aged 40 to 44
– Persons aged 45 to 49
– Persons aged 50 to 54
– Persons aged 55 to 59
– Persons aged 60 to 64 #
– Persons aged 65 to 74
– Persons aged ≥ 75

– Married persons
– Single persons
– Divorced persons
– Widowed persons
– Common-law #

EA Level Constraints

– Total households in EA
– Total persons in EA

Household WA Level Constraints

– Households of size 1 #
– Households of size 2
– Households of size 3
– Households of size 4
– Households of size 5

Appendix C – Additional Information on Statistics Used in Sampling Bias Study

Let X represent the known value for a 2A characteristic at the census division (CD) level and let $\hat{X}^{(0)}$ represent the Horvitz-Thompson estimator of X. $\hat{X}^{(0)}$ was calculated by multiplying the unweighted sample total for the characteristic of each sampled EA by the inverse of the realized household sampling fraction for the EA, and then summing the results to the census division (CD) level. Non-sampled enumeration areas (EAs) were excluded from the analysis. The standard deviation of $\hat{X}^{(0)}$, std $\hat{X}^{(0)} = \sqrt{V(\hat{X}^{(0)})}$ was calculated under the assumption that simple random samples of households were drawn independently in each EA (in fact, independent systematic random samples were drawn). Consequently, the variances were calculated at the EA level and summed to the CD level. The population S^2 values were used in the variance calculations. See Cochran, 1977, pp. 23-24, for variance formulas for person and family characteristics, and pp. 50-52, for variance formulas for household and dwelling characteristics.

Since the $\hat{X}^{(0)}$ values are Horvitz-Thompson estimators, they are unbiased for X. Sampling was done independently in different EAs. Therefore, the $\hat{X}^{(0)}$ values are the sum of *n* independent random variables, where *n* is the number of EAs in the CD. Since 90% of the CDs had more than 25 EAs with an average of 140, *n* is quite large in most CDs. Thus, according to the central limit theorem, $Z^{(0)} = (\hat{X}^{(0)} - X)/\text{std}(\hat{X}^{(0)})$ should follow an approximately normal (0.1) distribution (see Kendall and Stuart, 1963, p. 193). This, however, would not be the case if 2B responses were significantly biased for any reason.

The $Z^{(0)}$ values were produced for all 281 sampled CDs in Canada, for the 2A characteristics given in Chapter V. In order to evaluate the normality of the $Z^{(0)}$ values at the CD level, histograms of the $Z^{(0)}$ values overlaid with a normal PDF (Probability Density Function) were produced.

In addition, to test whether $Z^{(0)}$ was being selected from a normal distribution whose mean is zero (i.e. the sample selection procedure was unbiased), the mean $\overline{Z}^{(0)} = \sum_{i=1}^{m} Z_i^{(0)}/m$ was calculated where m = 281 (the number of CDs) and $Z_i^{(0)}$ is the value of $Z^{(0)}$ for the i^{th} CD. In addition, the standard deviation of the $Z_i^{(0)}$ was determined where $\text{std}^2(Z^{(0)}) = \sum_{i=1}^{m} (Z_i^{(0)} - \overline{Z}^{(0)})/(m-1)$. Then the T statistic $T_z = \sqrt{m}\,\overline{Z}^{(0)}/\text{std}(Z^{(0)})$ was calculated. If the sample selection procedure was unbiased, T should follow Student's *t* distribution with m-1 degrees of freedom. The probability of $|T_Z| > 1.960$ if the sample selection procedure was unbiased is less than 0.05. Thus, if $|T_Z| > 1.960$, the hypothesis that the sample selection procedure was unbiased will be rejected and the difference between the sample estimate and the population count will be said to be statistically significant at the 5% level.

Appendix D – 1986 Standard Error Adjustment Factors at National or Provincial Level and Percentiles of Weighting Area Level Factors

Characteristics	National or Provincial Factor	Percentiles of WA Level Factors					
		1	50	75	90	95	99
Population Characteristics							
Age							
Age groups 0-4, 5-9, 10-14, 15-19, 20-24, 25-29	0.18	0.05	0.19	0.29	0.35	0.49	0.52
Age groups 30-34, 35-44, 45-54, 55-59, 60-64, 5+, 15+	0.36	0.13	0.33	0.46	0.51	0.56	0.61
Age group 65+	0.00	-	-	-	-	-	-
Sex	0.00	-	-	-	-	-	-
Marital status							
Single, married (excluding separated)	0.25	0.04	0.23	0.31	0.42	0.49	0.55
Separated, divorced, widowed	0.88	0.55	0.84	0.98	1.06	1.15	1.20
Highest level of schooling							
Highest degree, certificate or diploma/total years of schooling	0.90	0.75	0.95	1.06	1.14	1.19	1.25
Major field of study	1.20	0.84	1.16	1.22	1.28	1.35	1.43
Mobility status							
Non-movers	1.21	0.83	1.23	1.27	1.32	1.36	1.41
Movers (migrants, non-migrants)	1.61	0.90	1.60	1.75	1.85	1.97	2.09
Period of immigration							
Before 1946, 1946-1966	0.98	0.76	1.02	1.10	1.22	1.37	1.45
1967-1977, 1978-1982, 1983-1986	1.51	0.80	1.45	1.55	1.78	1.90	2.11
Age at immigration	1.10	0.71	1.15	1.29	1.38	1.44	1.54

Characteristics	National or Provincial Factor	Percentiles of WA Level Factors					
		1	50	75	90	95	99
Population Characteristics							
Place of birth							
Born in Canada	1.09	0.82	1.08	1.16	1.18	1.20	1.21
Born outside Canada	1.35	1.11	1.34	1.43	1.60	1.67	1.75
Immigrant/Non-immigrant population	1.12	0.81	1.10	1.24	1.38	1.46	1.52
Citizenship							
Canada, by birth	1.13	0.88	1.14	1.17	1.20	1.27	1.32
Other	1.59	1.04	1.40	1.65	1.88	1.95	2.12
Ethnic origin							
English, French	1.20	0.73	1.16	1.25	1.31	1.40	1.46
Other	1.65	1.07	1.57	1.70	1.89	1.99	2.11
Home language							
English, French, English and French, English and non-official languages	1.12	0.50	1.09	1.35	1.75	1.89	2.09
Other language groups	1.76	0.99	1.68	1.89	2.01	2.20	2.41
Official language							
English, French, English and French	1.05	0.69	1.01	1.18	1.31	1.42	1.58
Other language groups	1.49	0.90	1.50	1.68	1.76	1.79	1.91
Mother tongue – English	1.21	0.83	1.23	1.27	1.32	1.36	1.41
Newfoundland, Prince Edward Island, Nova Scotia, British Columbia	0.92	0.24	0.96	1.45	1.62	1.90	2.23
Quebec	1.15	0.18	1.10	1.51	1.76	1.81	1.99
Other provinces	0.45	0.12	0.48	0.71	0.96	1.12	1.38
Canada	0.53	-	-	-	-	-	-
Mother tongue – French							
Quebec	0.42	0.14	0.45	0.52	0.61	0.76	0.91
New Brunswick	0.75	0.19	0.79	0.98	1.24	1.60	1.84
Other provinces	1.04	0.09	1.12	1.49	1.71	1.89	2.06
Canada	0.77	-	-	-	-	-	-

Characteristics	National or Provincial Factor	Percentiles of WA Level Factors					
		1	50	75	90	95	99
Population Characteristics							
Mother tongue – Other language groups	1.70	0.73	1.63	2.11	2.44	2.51	2.60
Industry/Occupation	0.92	0.25	0.80	1.13	1.25	1.31	1.38
Work activity in 1985	0.89	0.62	0.92	1.14	1.22	1.29	1.31
Weeks worked in 1985	0.94	0.68	0.99	1.18	1.29	1.33	1.39
Hours worked in reference week	0.83	0.63	0.85	1.01	1.14	1.19	1.24
Year last worked							
In 1986, in 1985, before 1985	0.89	0.60	0.94	0.99	1.05	1.11	1.20
Never worked	1.18	0.80	1.15	1.34	1.43	1.50	1.67
Class of worker							
Paid workers	0.72	0.56	0.75	0.86	0.93	0.95	0.98
Self-employed, unincorporated, unpaid family workers	0.93	0.68	0.96	1.08	1.13	1.15	1.18
Labour force status participation							
Employed	0.75	0.59	0.76	0.83	0.86	0.91	0.93
Unemployed	1.06	0.76	1.04	1.14	1.20	1.27	1.38
Not in labour force	1.25	0.91	1.30	1.43	1.50	1.58	1.63
Major source of income							
Wages and salaries	0.65	0.42	0.67	0.80	0.85	0.87	0.92
Other	1.05	0.71	1.00	1.12	1.17	1.20	1.24
Major source of income							
Wages and salaries	0.65	0.42	0.67	0.80	0.85	0.87	0.92
Other	1.05	0.71	1.00	1.12	1.17	1.20	1.24

Characteristics	National or Provincial Factor	Percentiles of WA Level Factors					
		1	50	75	90	95	99
Population Characteristics							
Disability							
Limited at home, school and work	0.94	0.69	0.96	1.11	1.29	1.34	1.42
Not limited	0.61	0.41	0.58	0.69	0.74	0.78	0.81
Census family status							
Husband, wife, child	0.20	0.05	0.20	0.24	0.26	0.28	0.31
Female lone parent	0.45	0.14	0.43	0.51	0.55	0.61	0.68
Male lone parent, non-member of a census family	0.68	0.35	0.65	0.79	0.89	0.99	1.14
Economic family status							
Husband, wife	0.14	0.06	0.16	0.21	0.28	0.34	0.36
Lone parent, child	0.32	0.16	0.34	0.39	0.44	0.47	0.53
Other family members	0.74	0.24	0.70	0.84	1.03	1.09	1.18
Number of persons in the census family	0.04	0.00	0.00	0.05	0.07	0.09	0.11
Number of persons in the economic family	0.18	0.08	0.19	0.24	0.33	0.41	0.45
Age of husband, wife or reference person of economic family	1.42	0.80	1.37	1.53	1.60	1.78	1.91
All other population characteristics	1.00	-	-	-	-	-	-
Household and Dwelling Characteristics							
Structural type							
Single detached	0.33	0.05	0.35	0.55	0.67	0.75	0.89
Apartment less than 5 storeys	0.57	0.12	0.56	0.70	0.83	0.99	1.26
Other	0.91	0.18	0.88	0.99	1.18	1.23	1.32

Characteristics	National or Provincial Factor	Percentiles of WA Level Factors					
		1	50	75	90	95	99
Household and Dwelling Characteristics							
Tenure	0.00	-	-	-	-	-	-
Period of construction	0.78	0.61	0.75	0.82	0.89	0.99	1.24
Main type of heating equipment/principal heating fuel	0.87	0.18	0.86	1.04	1.12	1.25	1.32
Central heating equipment							
With	0.42	0.09	0.38	0.54	0.60	0.70	0.89
Without	0.78	0.23	0.79	0.91	1.03	1.12	1.20
Household size							
One-person household	0.00	-	-	-	-	-	-
Other	0.76	0.19	0.72	1.09	1.17	1.21	1.30
Number of rooms	0.80	0.57	0.78	0.90	0.97	1.10	1.20
Age of household maintainer							
25-34, 55-64, 65-74, 75+	0.25	0.06	0.24	0.35	0.48	0.53	0.62
0-24, 35-44, 45-54	0.92	0.38	0.90	1.05	1.14	1.21	1.30
Sex of household maintainer							
Male	0.20	0.09	0.24	0.31	0.34	0.36	0.37
Female	0.47	0.16	0.43	0.54	0.64	0.74	0.89
Gross rent/gross rent as a percentage of household income	0.75	0.48	0.79	0.91	0.94	0.96	1.01
Owner's major payments/owner's major payments as a percentage of household income	0.84	0.62	0.87	0.95	1.01	1.04	1.11

Characteristics	National or Provincial Factor	Percentiles of WA Level Factors					
		1	50	75	90	95	99
Household and Dwelling Characteristics							
Household income	0.75	0.51	0.73	0.82	0.90	0.95	1.03
Value of dwelling	0.90	0.67	0.91	1.00	1.05	1.12	1.18
Registered condominium							
Part	0.63	0.18	0.59	0.84	0.93	1.11	1.30
Not part	0.15	0.07	0.14	0.19	0.28	0.39	0.47
Household type – One-family households							
Without additional persons	0.22	0.05	0.20	0.27	0.33	0.36	0.40
With additional persons	0.50	0.20	0.48	0.61	0.72	0.74	0.79
Household type – Non-family households	0.00	-	-	-	-	-	-
Household type – Other	1.12	0.54	1.05	1.26	1.40	1.51	1.67
All other household and dwelling characteristics	1.00	-	-	-	-	-	-
Census Family Characteristics							
Census family structure							
Husband and wife	0.20	0.09	0.21	0.26	0.29	0.33	0.36
Male lone parent	0.64	0.21	0.62	0.81	0.84	0.91	1.04
Female lone parent	0.46	0.19	0.45	0.57	0.65	0.69	0.74
Census family type							
Primary family	0.23	0.04	0.24	0.28	0.31	0.34	0.39
Secondary family	0.90	0.62	0.93	1.15	1.28	1.33	1.40
Age groups of children at home	0.78	0.40	0.70	0.91	0.98	1.09	1.19

Characteristics	National or Provincial Factor	Percentiles of WA Level Factors

		1	50	75	90	95	99
Census Family Characteristics							
Labour force activity of husband, wife or lone parent							
Husband, lone parent, husband and wife in labour force	0.40	0.23	0.43	0.50	0.55	0.59	0.71
Wife in labour force	0.61	0.41	0.60	0.68	0.74	0.78	0.82
Other	0.72	0.30	0.68	0.80	0.90	0.99	1.12
Work activity in 1985 of husband, wife or lone parent							
Worked in 1985	0.48	0.11	0.45	0.50	0.54	0.57	0.59
Did not work in 1985	0.93	0.60	0.90	1.04	1.18	1.26	1.30
All other census family characteristics	1.00	-	-	-	-	-	-
Economic Family Characteristics							
Economic family structure							
Husband and wife families	0.29	0.13	0.30	0.36	0.48	0.56	0.68
Non-husband and wife families	0.56	0.35	0.50	0.66	0.81	0.90	1.06
Mother tongue of family reference person – English							
Newfoundland, Prince Edward Island, British Columbia	0.25	0.09	0.20	0.31	0.45	0.66	0.91
Quebec	0.49	0.25	0.47	0.50	0.69	0.83	1.05
Other provinces	0.18	0.07	0.19	0.22	0.24	0.27	0.31
Canada	0.27	-	-	-	-	-	-
Mother tongue of family reference person – English							
Newfoundland, Prince Edward Island, British Columbia	0.25	0.09	0.20	0.31	0.45	0.66	0.91
Quebec	0.49	0.25	0.47	0.50	0.69	0.83	1.05
Other provinces	0.18	0.07	0.19	0.22	0.24	0.27	0.31
Canada	0.27	-	-	-	-	-	-

Characteristics	National or Provincial Factor	Percentiles of WA Level Factors					
		1	50	75	90	95	99
Economic Family Characteristics							
Mother tongue of family reference person – Other than English or French							
Newfoundland, Nova Scotia	0.75	0.38	0.74	0.80	0.91	0.99	1.10
Other provinces	0.50	0.21	0.45	0.57	0.82	0.84	0.99
Canada	0.56	-	-	-	-	-	-
All other economic family characteristics	1.00	-	-	-	-	-	-

Appendix E – Products and Services

Packaging census data so they are meaningful and accessible to clients, whether they are government decision-makers, policy analysts, librarians, marketing specialists, researchers, students, etc., is the key to ensuring that the value of the data is maximized. There are several new product and service features for 1996.

1. Increased Accessibility Through Electronic Media

More clients asked for census materials to be available in electronic formats which can be used with personal computers. While some key printed products have been retained, more census data were produced on CD-ROM and on diskette. These formats contained Windows-based presentation and tabulation softwares to make the data easy to use. For the first time, clients were able to obtain information free of charge on the Internet through the Statistics Canada's Web site: http://www.statcan.ca.

2. Small Area Data Available Sooner

Census data at smaller levels of geography were made available much sooner than in previous years. On each release day, profile data were available for areas at the community levels (census subdivisions and census divisions) and, one month after their release, data for areas as small as census tracts, enumeration areas and forward sortation areas.

3. Census Tabulations Available by Postal Code

As part of the standard product line, basic summary tabulations and area profiles were available for forward sortation areas, which represent the first three characters of the postal code. Data for the full postal code can be obtained as a custom service, subject to confidentiality restrictions.

4. New Information Collected in 1996

For the first time, data will be published for unpaid household activities, place of work for all levels of geography, mode of transportation to work and population groups.

5. Improvement of Geography Products

Not only has the quality of many of the maps used for the release of census data been improved, a map series on federal electoral districts has been reintroduced. *GeoSuite* (formerly GeoRef), the Windows-based electronic tool which allows clients to explore the links between different levels of geography, has also been improved with the addition of enumeration area reference lists.

Bibliography

Census Operations Division produced the following portions of this report: Introduction, Appendices A and E, and Regional Reference Centres.

Béland, Y. "Results and Methodology of the 1986 Sampling Variance Study". Statistics Canada Internal Report, 1990.

Bankier, M., S. Rathwell, and M. Majkowski. "Two-step Generalized Least Squares Estimation in the 1991 Canadian Census". Methodology Branch Working Paper, Social Survey Methods Division, 1992. Catalogue number 92-007E.

Bankier, M., A.-M. Houle, and M. Luc. Calibration Estimation in the 1991 and 1996 Canadian Censuses. Proceedings of the Survey Methods Research Section, American Statistical Association, pp. 66-75. 1997.

Brackstone, G.J. and J.N.K. Rao. "An Investigation of Raking Ratio Estimators". Sankhya, Volume 41, Series C, Pt. 2, p. 97-114. 1979.

Cochran, W. Sampling Techniques 3rd Edition. John Wiley and Sons: Toronto, 1977.

Dominion Bureau of Statistics. Eighth Census of Canada, 1941, Administrative Report of the Dominion Statistician, Ottawa: King's Printer, 1945.

Dominion Bureau of Statistics. Ninth Census of Canada, 1951, Vol. XI, Administrative Report, Ottawa: Queen's Printer, 1955.

Dominion Bureau of Statistics. Sampling in the Census. S.M.S.03.5, 1968.

Dominion Bureau of Statistics. 1961 Census of Canada. General Review, Bulletin 7.2-12, Ottawa: Queen's Printer, 1970. Catalogue number 99-537.

Fellegi, I.P. "Response Variance and its Estimation". Journal of the American Statistical Association, 59, pp. 1016-1041. 1964.

Hansen, M.H., W.N. Hurwitz, and M.A. Bershad. "Measurement Errors in Censuses and Surveys". Bulletin of the International Statistical Institute, 38, pp. 359-374. 1959.

Kendall, M.G. and A. Stuart. "The Advanced Theory of Statistics", Volume 1, Charles Griffin and Company Limited, London, 1963.

Kruszynski, G. Evaluation of the 1996 Weighting Areas. Internal Report, Geography Division, Statistics Canada, 1999.

Majkowski, M. “1991 Census 2A/2B Discrepancies”. Statistics Canada Internal Report, 1992a.

Majkowski, M. “Investigation into Large Population/Estimate Differences in the 1991 Census”. Statistics Canada Report, 1994.

Royce, D. “The Use of Sampling in the 1981 Canadian Census”. Statistics Canada Internal Report, 1983.

Sarndal, C., B. Swensson, and J. Wretman. Model Assisted Survey Sampling. Springer-Verlag: New York, 1992.

Statistics Canada. 1971 Census of Canada. General Review, Vol. VI, Part 1, Ottawa, 1976. Catalogue number 99-740.

Statistics Canada. 1976 Census of Canada. Administrative Report, Part 1, Ottawa, 1980. Catalogue number 99-850.

Statistics Canada. 1976 Census of Canada. Quality of Data, Series 1: Sources of Error - Sampling and Weighting. Ottawa, 1980. Catalogue number 99-844.

Statistics Canada. 1981 Census of Canada. Summary Guide: Total Population. Ottawa, 1983. Catalogue number 99-902.

Statistics Canada. 1986 Census of Canada. Census Handbook. Ottawa, 1988. Catalogue number 99-104E.

Statistics Canada. User’s Guide to the Quality of 1986 Census Data: Sampling and Weighting. Ottawa, 1990. Catalogue number 99-136E.

Statistics Canada. 1991 Census of Canada. Census Handbook. Ottawa, 1992. Catalogue number 92-305E.

Thivierge, S. Bias in the 1996 Census Sample. Internal Report, Social Survey Methods Division, Statistics Canada, 1999.