

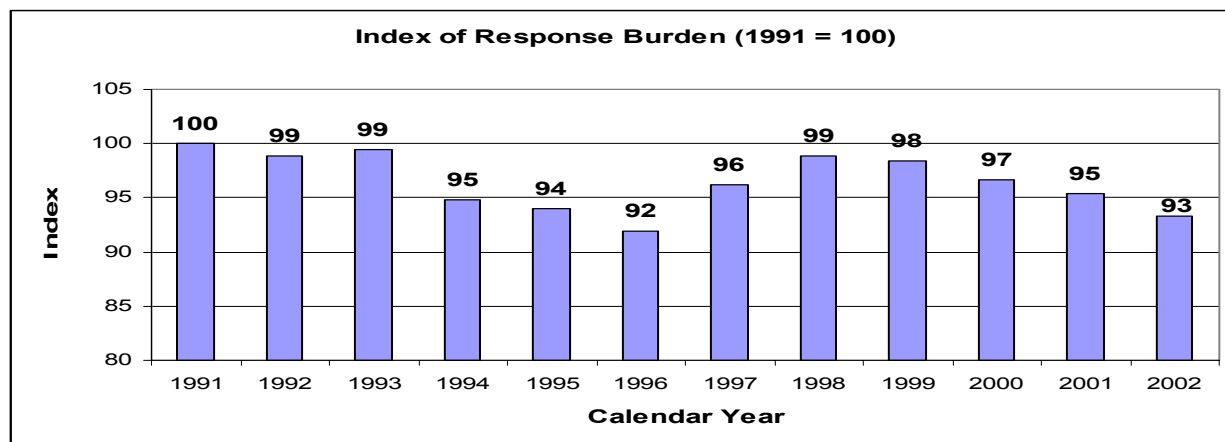
Costs and Response Burden Reductions at Statistics Canada

Prepared by
Bernard Lefrançois, Senior Analyst¹
Industry Measures and Analysis Division
Statistics Canada

July 18, 2003

Introduction

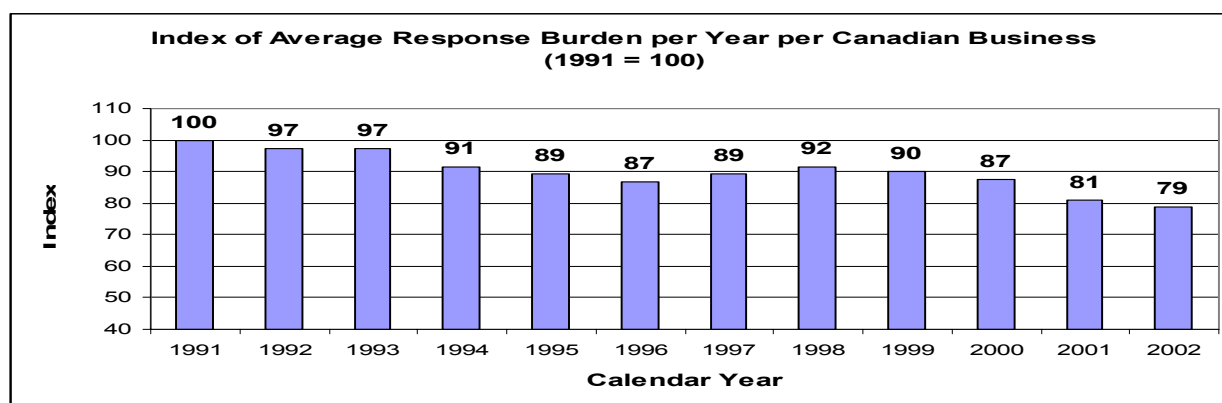
Statistics Canada continuously reviews its programs and practices in order to reduce costs and the burden imposed on our respondents. Our performance on both counts is reported annually to Parliament for public scrutiny.² Statistics Canada has measured the burden it imposes on Canadian businesses for years, but only on a regular annual basis since 1991. While there are many interpretations of what constitutes response burden, Statistics Canada has chosen to measure burden in terms of the cost, in hours, we impose on the business community to complete our various surveys. We refer to this as “compliance cost”. The response burden is calculated using the survey frequency, its sample size, the average time required to complete the survey, and the response rate. The response burden is calculated separately for large and small respondents, with the separation being \$2 million in gross business income.



The burden of response for businesses now stands at about 33% of the 1978 levels. Response burden declined in the mid 1990s primarily as a result of the redesign of the monthly Survey of Employment, Payrolls and Hours to use administrative data. While burden increased in the late 1990s with the introduction of a more comprehensive annual survey program, known as the Unified Enterprise Survey, progress has been made largely as a result of reductions in time required by respondents to complete surveys (i.e. content reduction and simplification) and increased use of administrative data. Total hours of burden declined approximately two percent in 2002 compared to 2001. Even though overall response burden on businesses declined only seven percent since 1991, the Canadian economy has expanded appreciably over the same period. Thus in 2002 the average response burden per business is twenty-one percent lower than in 1991.

¹ The author gratefully acknowledges the comments received from many colleagues at Statistics Canada.

² See Statistics Canada (2002).



Monitoring our costs and the response burden shows our progress and guides us in the identification of initiatives to reduce both further. Examples of such initiatives are the removal of a number of questions from ongoing surveys, the elimination of surveys whose usefulness has lapsed, and adhering to general principles in our search for more efficient and less burdensome business survey techniques. One such principle is a common approach to business survey designs. Across surveys of specific industries or business characteristics (such as employment) a common basic sampling plan is used: the samples are stratified by industry, regions, and size of establishments, with only minor adaptations for specific surveys. Other principles are the use of thresholds to limit the number of small establishments surveyed,³ the use of a unique Business Register as survey frame as well as employing standardised software tools for survey processing.

More recently, three new major initiatives have been launched to reduce costs and the response burden, particularly for businesses. The first two concern our practices: we have introduced a specific program to interact with the largest enterprises operating in Canada, known as the Key Provider Managers Program, and we have implemented Electronic Data Reporting facilities. The third represents a strategic streamlining initiative that consists of using tax data⁴ for filling data gaps and for sample replacement, particularly to obtain the relevant information from those small establishments excluded by the use of thresholds in surveys.

In this paper we will describe the above initiatives. First, we will describe the Key Provider Managers Program that has been specifically introduced to facilitate our relationships with the largest enterprises in Canada in order to reduce their response burden and to obtain the appropriate information that we seek. Second, we will briefly describe the work in progress toward the generalised use of Electronic Data Reporting through the Internet. Both these programs aim to reduce the burden of response by changing our data collection practices.⁵

Next we will describe how the tax data strategic streamlining initiative has been applied to three different programs. For the first two of these, the monthly Survey of Employment, Payrolls and Hours, and the Monthly Restaurants, Caterers and Taverns Survey, tax data are used to complement the information obtained from surveys. The third and last program we describe is the redesign of the Monthly Wholesale and Retail Trade Survey which combines the elements discussed above in order to reduce its costs and the response burden. Together, these three monthly surveys provide the bulk of the information used to derive the monthly estimates of real valued added for the Service sector in Canada.

Note however that new practices and processes are never adopted before a thorough analysis of their impact on the quality of the statistical information produced. In addition, the introduction of new survey

³ See Statistics Canada (1998, 1999a).

⁴ For a review of tax data sources in Canada see Bissett (1999).

⁵ A more comprehensive paper describing Statistics Canada's relations with its business respondents will be presented at the forthcoming Australian Bureau of Statistics – Statistics Canada Management Meeting slated for September 24-26, 2003; see Baxter (2003 forthcoming).

instruments, or changes to current ones, such as questionnaires, the use of Computer Assisted Interviewing, and Electronic Data Reporting, are generally preceded by pilot tests. These pilot tests allow us to study the feasibility of changes to collection methods as well as their potential impact on responses (mode effects), and they help us identify and adjust for difficulties in their large-scale implementation. The pilot tests thus ensure us that the changes will preserve the goodwill of our respondents and the quality of the results obtained.

The Key Provider Managers Program⁶

In 1997 Statistics Canada undertook a major redesign of its entire framework for conducting annual business surveys. The goals of this redesign were to improve provincial economic statistics and to fill some gaps in industry coverage, especially in the service sector, as well as to standardise the common features of business surveys, while providing us with a better management of the response burden.⁷ The resulting program is known as the Unified Enterprise Survey (UES). A major element of the UES is the Key Provider Managers Program that was modelled on a similar program at the Australian Bureau of Statistics.

Of the hundreds of thousand of enterprises in Canada, very few are large and complex, i.e. whose activities span more than one province or industry. The top 300 enterprises in Canada represent about one-third of the country's output in terms of gross business revenues. They are thus very important for many of our business surveys. Incomplete or inaccurate reporting on their part could significantly impact the quality and accuracy of the statistical estimates produced. They are therefore generally included in every survey for which they are part of the universe.

The Key Provider Managers Program was put in place specifically to enhance Statistics Canada's working relationships with large and complex enterprises, and particularly, though not exclusively, the 300 largest. It consists in dedicating some employees, known as Key Provider Managers (KPM), to serve as communication channels between these key data providers and Statistics Canada for addressing the issues and concerns of both. Other fundamental objectives of the KPM Program are to obtain a clear understanding of the structure of these complex enterprises, and to improve response in terms of timeliness, completeness and accuracy, as well as the quality and coherence of the data.

Each KPM manages a portfolio of enterprises and is responsible for identifying counterpart focal points within the target enterprises. With these enterprise focal points, the role of the KPM is to build tailored relationships for articulating and rationalising Statistics Canada's data requirements; negotiating reporting arrangements; analysing data coherence; and for managing the response burden and the many issues that affect the quality and timeliness of business survey reporting for large and complex enterprises.

The KPM seeks out a single focal point in the enterprise, preferably at a senior level within the organisation: a Controller, Vice President or Director of External Reporting. The enterprise focal point is often influential in resolving reporting problems throughout the enterprise. The enterprise focal point however, cannot always co-ordinate the survey activity within these large organisations. The KPM often requires multiple contacts within each enterprise in order to facilitate the collection of survey data. However, time and again it is the senior contact and his influence in the organisation that provides results. All survey respondents in an organisation should be aware that the KPM can assist them with any survey-related issues.

In meetings with the enterprise to discuss burden or reporting difficulties, a KPM looks at the issues and priorities the respondent faces and if required, negotiates reporting arrangements that respect that reality. The goal is clear, however. The KPM wants to build long term reporting arrangements that are both workable for the survey and agreed to by the respondent.

⁶ Much of this section is an abbreviated version of Gaudreau & Hughes (2000) where more information on Statistics Canada's experience with the KPM program can be found.

⁷ See Statistics Canada (1996, 1999b) and Tourigny, Pursey & Whitridge (2001).

Response burden can be described and perceived in many ways and inevitably has a direct impact on the quality of the data reported (or not) by the respondent. An inventory of surveys, along with the coverage, contacts and reporting history, is critical to the KPM's initial discussions with an enterprise and to the assessment and management of the overall response burden. In most cases, once the KPM had pulled together the inventory of surveys, there was no doubt that the reporting burden on these enterprises was significant. The KPM approach helped to minimise this burden, negotiating toward a win/win outcome for both the enterprise and the survey managers.

As the KPM knowledge of the enterprise increases, so do the opportunities to ease the reporting burden for these respondents. By obtaining an understanding of how the financial and production data are maintained by these enterprises, the KPM can streamline their individual reporting arrangements and look at alternative reporting methods such as utilising existing reports and alternative formats to reduce burden. Tailored reporting arrangements are required if we expect to accurately measure economic activity while managing response burden. KPMs must customise the Agency's generic survey-taking to suit the enterprise operations and financial reporting structure as well as adjusting data collection and follow-up to avoid peak financial and planning cycles. In practice, KPMs do not disturb existing reporting arrangements or interfere with regular collection processes except when requested to do so by either the enterprise or the survey managers. The KPM has become the point of contact for survey managers when they want answers or need problems resolved quickly, rather than a barrier to their relationship with the enterprises.

For coherence purposes the data collected from the enterprise and its establishments must be comprehensive, unduplicated, and fit together as a single, inter-related set of information consistent with the consolidated income statement and balance sheet data provided for the enterprise as a whole. Coherence analysis is an important aspect of the KPM role at Statistics Canada.

The introduction of the KPM program is a major change in the way Statistics Canada communicates with its most important business respondents. The long-established tradition was that each survey program managed its own respondent community, with little co-ordination among surveys. Moving from this decentralised approach to the highly co-ordinated KPM model was a balancing act between keeping individual surveys going smoothly while at the same time dealing with large enterprises in a more coherent and fair manner.

The KPM program has focused efforts to improve data quality on the survey population that is most important to the Agency's business estimates. By developing relationships with some of the largest enterprises in Canada, the program has made significant strides in improving response rates, data quality and timeliness with these respondents. Senior management within these enterprises appreciates the Agency's efforts toward rationalising and integrating collection activities. They view the KPM Program as an example of the government's attitude towards businesses and are encouraged by the sense of co-operation, understanding, flexibility and assistance provided by the KPM. Focal points like to be able to put a face to Statistics Canada and to know they have an advocate in their dealings with the Agency. Time and again, they were gratified to finally meet in person someone representing Statistics Canada's viewpoint and were extremely pleased with the single focal point approach.

By opening the lines of communication we have listened to respondents and demonstrated our commitment to take action to address their concerns. The KPM experience has taught us a great deal about improving the way we do business, not only with the most significant enterprises, but also with all our data providers. Because of the success of the KPM program, we can see it broadened to multinational enterprises. Statistics Canada recently proposed to set up an international project where national statistical offices (NSOs) would cooperate in establishing an experimental process to obtain insights into the problems associated with measuring the activities of multinational enterprises. One element of this proposal calls for participating NSOs to designate co-ordinators to manage relations with these enterprises. For more information on this proposal, see Barnabé (2003).

Electronic Data Reporting⁸

Canada is one of the most Internet aware countries in the world. In 2001, more than 5.8 million households (49% of all households in Canada) had regularly used the Internet at home.⁹ In 2002, 76% of businesses accounting for 97% of economic activity in Canada used the Internet.¹⁰ In addition, the Canadian government has launched in 2000 an important initiative known as Government On-Line whose objective is to make all government services available to Canadians via an online channel by the end of 2005.¹¹

Statistics Canada recognises that there is already a large proportion of the Canadian population and businesses that would prefer the option of filing their survey and Census returns electronically. A recent survey indicated that 56% of Canadians and 85% of businesses want to use the Internet to report to surveys. Our own research indicates that over three-quarter of respondents with an access to the Internet would use Electronic Data Reporting (EDR) if it were available. In 2002, nearly 40% of the personal tax returns in Canada were filed electronically. In fact, when consulted by the Canadian Information Office about their expectations regarding e-government, Canadians indicated that being able to respond on-line to government surveys is a top priority. Only electronic filing of tax returns ranked higher.

These factors, together with Statistics Canada's interest in offering more flexibility to respondents, are some of the factors that have led to the development of strategies for EDR during the last few years. Over the past decade, electronic data collection has evolved from the mailing of diskettes to the completion of web surveys. Our current strategies are still in evolution, and are expected to continue to be so as the technology is still evolving quickly, as are the perceptions and habits of their users. The principal constraint in the adoption of any specific technology for EDR as a survey instrument by Statistics Canada has been, still is, and will be our unbending commitment to protect the confidentiality and the privacy of our respondents. Currently we are focusing on two methods for receiving data electronically. Both methods are based on the most secure approach possible, which requires encryption on the respondent's workstation.

The first method of EDR is known as the Data Return Module (DRM) and is used for a number of economic and institutional surveys. It consists of a secured infrastructure to receive a file in any format; especially completed questionnaires developed by Statistics Canada with software respondents already have, e.g. Microsoft Excel spreadsheets designed to look like questionnaires. This has proven to be a good choice, as the respondents are already familiar with the software, and may even link or copy the information from their usual files to complete the survey. Edit rules can be available on the Excel form itself. To ensure the security of this kind of transmission respondents need to install software on their workstations. The software is a small module (approximately 2.4 MB) of tools for encrypting the data and sending back the files. In the coming months we expect to replace this with a web-upload that will ensure the security of the data without requiring anything to be resident on the respondent's computer.

The second method of EDR is a secured infrastructure for web-based questionnaires that respondents can complete over the Internet. The questionnaire consists of HTML pages with built-in Java scripts for editing, navigation and encryption. Respondents access the survey pages on Statistics Canada's web server using the password information we sent them with the URL link. This web server is known as the Secure Internet Response Site (SIRS).¹² No respondent data are transferred across the Internet while the survey is being completed. Once the survey is completed, the data are encrypted on the respondent's workstation and then sent through the Secure Socket Layer to our server. We have implemented a generalized eXtensible Mark-up Language (XML) model to structure the returned data. This gives us the required flexibility for easy implementation in multiple surveys.

⁸ Much of this section is an updated and abbreviated version of Mayda (2002). See also Essoltani & Zorzi (2003).

⁹ Statistics Canada's 2001 Household Internet Use Survey. See www.statcan.ca/Daily/English/020725/d020725a.htm.

¹⁰ Statistics Canada's 2002 Survey of Electronic Commerce and Technology. See www.statcan.ca/Daily/English/030402/d030402a.htm.

¹¹ See www.gol-ged.gc.ca/index_e.asp.

¹² See www.statcan.ca/sirs.

In the recent past, Statistics Canada has placed an emphasis on including EDR as a data collection option for several business and agricultural surveys. We started with pilot projects for a diverse range of important surveys involving respondents from households, universities, businesses, and federal departments. Pilot projects included the 2001 Census of Population, the Unified Enterprise Survey and the Business Payroll Survey. Both the DRM and web-based solutions have now been implemented with success in many collection operations. Currently, 11 monthly and 2 quarterly business or agriculture surveys offer an EDR option. In 2004 we will stage a full dress rehearsal of the EDR option that we plan to offer for the 2006 Census of Population. Some of the issues we have encountered while deploying EDR are briefly described next.

The Data Return Module (DRM) is generally used with a standardized questionnaire or file format. It is also used without a standardized questionnaire for several large companies who respond to many Statistics Canada surveys. These are the key data providers described in the preceding section. The DRM is a good approach for these respondents, as it is the most accommodating to them and reduces response burden. The DRM allows respondents to send whatever kind of file they have, whenever they can. The downside is that it is relatively complex to convert the information when it is returned. For this reason, we try to encourage standardized formats in the majority of cases.¹³

EDR must be user-friendly. Experience has taught us that if respondents are not successful the first time, they are reluctant to try again. Usability testing is an important part in ensuring that an application is intuitive. However, we have also found that a toll free help-line is also an essential component for the success of EDR from a respondent perspective. Requiring respondents to install special software on their workstations has proved to be a major deterrent, in addition to the inherent complexity related to the diverse environments (operating systems, browser versions and types, etc.) in which the software must operate. Now, the more recent versions of the most common web browsers provide adequate functionalities and levels of security so that no separate software is required.

Some of the advantages of the web-based option include the ability for questionnaire navigation, built-in skip patterns, and real-time edits. However, we are hesitant to program the same edits that we usually apply to responses in other methods in the fear that we might irritate the respondent if edits are triggered too often automatically. We did adopt an unwritten policy that no edits should be mandatory on the web-based form. That is, respondents always have the option to ignore an edit message and to submit their data. We have also found that the use of individualized passwords and confirmation codes upon receipt of the data serve to make the respondents more at ease with the security of their data transmission.

The organisation's ability to maintain the hardware, the software and its documentation, as well as the infrastructure involved in EDR should not be underestimated. Continual updates, changing products and volume of responses are factors that should not be ignored, as well as the costs related to them. A very important factor in the acceptance of EDR by survey managers is the way in which it can be integrated seamlessly into the usual collection process, as EDR is and will remain in almost all cases one more option for respondents. In addition, the integration of EDR in an existing survey program generally requires bridging software. In these circumstances, it is important to think of generic applications usable for different surveys. We have developed a standardised way to receive specifications from the client for EDR applications in terms of edits, navigation, look and feel, data capture and output file format. Plans to evaluate the impact the introduction of EDR may have on survey results are under way.

Most of our experience derives from business surveys. We have found that in many instances access to the Internet is available within a company, but not necessarily at the desktop of the respondent. There are also companies with firewalls that prevent the transmission of data. But at times confidentiality and security may appear to be more of an issue for us than for some of our business respondents, since we frequently receive data in unencrypted email attachments. The viewpoint of the largest enterprises in

¹³ The development of the eXtensible Business Reporting Language (XBRL), a common language to exchange financial information between various software packages, is progressing; see www.xbrl.org. Statistics Canada is following closely the progress of XBRL and intends to co-operate with the various participants to try to incorporate the functionality for survey reporting in the context of this emerging standard based on XML.

Canada is crystal-clear however. The Key Provider Managers inquired about their concerns regarding the use of EDR. The most cited concern was the need for a reliable security framework for the site. For the most part, large corporations have a general sense that their data are protected by Statistics Canada even if they may not be aware of the particular procedures in place to accomplish this. In circumstances when the information is especially sensitive, they may even be reluctant to divulge the information to Statistics Canada. So far the Key Provider Managers have been able to satisfy any respondent concerns, and the personalised approach gives additional credibility to these assurances.

Statistics Canada has embarked on a journey that will result in the majority of our business and agricultural surveys offering an EDR option. Our ultimate goal would be to have all our respondents use EDR. However, EDR is and will remain for the foreseeable future only one more option for respondents. It appears that certain types of surveys would not necessarily benefit from EDR; examples are surveys where the role of the interviewer is considered very important, and those with long and complex questionnaires. Offering an electronic option to survey respondents can not only improve the timeliness and quality of the data but gives respondents greater flexibility in choosing how and when they choose to complete a questionnaire.

EDR as a collection option is in its nascent stages and will continue to evolve. EDR has the potential to reduce response burden and generate cost savings over time if adopted by a sufficient number of respondents. In situations where the required information exists within the respondent systems in such a way that at the push of a button they may send it to us, we cannot hope to reduce their burden further. The biggest challenges related to EDR are not necessarily technological but consist of a better understanding of the issues related to security, risks or convenience by the respondents and by the statistical agency. Even if we develop EDR options for respondents, the success of the new approaches depends largely on their willingness to use these options instead of the other methods to respond to our surveys. The notion of measuring respondent burden is not only applicable to the time and effort to respond to a questionnaire but also has to include all those other aspects associated with the collection method. These aspects are sometimes less quantifiable but they are also very important especially when the respondents can choose their preferred collection method and when a statistical agency tries to implement widely a new one. If a respondent would ask “What’s in it for me?” we need to be able to show clearly the advantages of EDR from his perspective in order to increase the popularity of this collection method over time.

Redesign of the Survey of Employment, Payrolls and Hours¹⁴

The monthly Survey of Employment, Payrolls and Hours (SEPH) provides estimates of the levels and changes in payroll employment, paid hours and earnings. The data are compiled at detailed industrial levels for Canada, the provinces and the territories. The target population is composed of all employers in Canada, except those primarily involved in agriculture, fishing and trapping, private household services, religious organizations and military personnel of defence services.¹⁵ In all, there are close to one million establishments in scope for the survey.

SEPH was designed at the beginning of the 1980’s as a stratified sample of establishments.¹⁶ The estimation procedure was strictly based on the survey design weights and the data directly collected from the sampled establishments. That is, no auxiliary information was used in estimation. The sample size was about 70,000 establishments.

Enterprises in Canada are required to remit to Canada Customs and Revenue Agency (CCRA) certain amounts deducted from their employees’ salaries and wages. Concerns with response burden led Statistics Canada and CCRA to investigate whether payroll employment and gross monthly payroll information could be added to the monthly payroll deduction remittance form. Since these two variables were already

¹⁴ The main references for SEPH and its redesign are: Grondin & Lavallée (2002), Leduc (2001), Rancourt & Hidirolou (1998) and Statistics Canada (2001). See also www.statcan.ca/english/sdds/2612.htm.

¹⁵ Note that SEPH produces estimates for general government services at the provincial and federal levels from information provided by Public Institution Division.

¹⁶ See Schiopu-Kratina & Srinath (1991).

collected by SEPH, those same variables collected by CCRA instead, would constitute an excellent source of auxiliary information, and there would be a net reduction in total response burden. After focus group testing to gauge the reaction of potential respondents revealed that these changes to the remittance form were not perceived to increase significantly the response burden, the questions were added to the form beginning in January 1993. Several studies were carried out to evaluate this new information and it was concluded that SEPH would greatly benefit from using it.¹⁷

Hence, SEPH was one of the first sub-annual business surveys to incorporate administrative data to supplement survey results in recent years. The use of administrative data contributes to the reduction of survey costs and response burden. They also enhance the overall estimates of SEPH payroll employment and increase our ability to produce small-area estimates. These improvements are due to the timeliness of administrative data and to a better coverage of the surveyed population.

A unique feature of the SEPH redesign was its three-phase implementation. We adopted this approach because the inclusion of the two additional questions on the remittance form was itself staged, and we had to let respondents get accustomed to answering them. Statistics Canada also needed to familiarise itself with the quality of this new data source, and it allowed us to amortise the costs of the redesign over many years.

The three phases of the redesign consisted in gradually including the administrative data in lieu of survey data to estimate employment and monthly payroll. The administrative data were supplemented with a small survey of establishments, the Business Payrolls Survey (BPS), to estimate variables not available from the administrative source, such as the weekly component of payroll, the employment category (salaried, paid by the hour and others such as commission workers) and the hours paid. The estimates derived from the administrative source are combined with the results of the BPS to produce estimates of the full range of SEPH variables.¹⁸

Phase I of the redesign was implemented with the March 1994 reference month when administrative records were used for small employers (less than 100 employees). They accounted for about 30% of total employment in Canada. As a result, the SEPH monthly sample size was reduced by 30,000 respondents, from an original size of 70,000. At the same time a sample of 2,500 establishments was drawn for the Business Payrolls Survey (BPS).

In May 1996, the use of administrative data was extended to cover medium (less than 300 employees) and large businesses that operated in unique combination of industrial activity and province. By then the businesses included in the administrative portion of SEPH represented about 70 % of the total employment in Canada. The implementation of this Phase II resulted in a further reduction of some 24,000 respondents. The BPS sample was increased to 6,000 establishments.

The last phase of the redesign (Phase III) was implemented in May 1998. It consisted in extending the use of administrative data to businesses engaged in multiple industrial activities or in different provinces. This third and last phase of the redesign allowed a further reduction of 16,000 establishments in the combined monthly sample size, leaving only a sample of 10,000 establishments included in the BPS.

With the January 2001 data, SEPH migrated to the new North American Industrial Classification System (NAICS). The increase in required detail necessitated a slight increase to the BPS sample size, which now stands at 11,000 establishments. With the completion of the survey redesign, SEPH's monthly sample size was reduced from 70,000 to 11,000 establishments. This represents a reduction of 708,000 contacts on an annual basis. Note that a rotation of approximately one twelfth of the sampled establishments occurs each month.

¹⁷ See Lee & Croal (1991) and St-Martin (1992).

¹⁸ SEPH produces estimates for eleven base variables from which all other variables are derived.

In SEPH, administrative data on payroll deductions are used in a variety of ways. They are used to update the Business Register, the frame for the BPS, and to evaluate the quality of the estimates produced from the BPS. They are also used as auxiliary information to estimate through regression total hours worked and summarised earnings, i.e. total earnings including overtime and special lump-sum payments for the last seven days of the reference month. Quality monitoring is performed at both the macro and micro levels. At the macro level, the administrative employment and payrolls totals provide a convenient source of information to assess the quality of the corresponding estimates produced from the BPS. At the micro level, a number of studies could be performed during the redesign, and they could still be carried out on a regular basis to monitor the quality of the BPS sample.

The Business Payrolls Survey (BPS) provides the basis for the estimation of SEPH variables not available from the administrative source such as: the weekly component of the gross monthly payrolls, the total number of paid hours (regular hours and overtime) and the allocation of hours, earnings and employment for three categories of employees (salaried, paid by the hour and others such as commission workers). Allocation of the BPS sample is designed to achieve targeted coefficients of variation at the provincial and industrial levels in order to produce reliable estimates of regression coefficients and ratios at the model group level. Model groups form the level at which regression and ratio estimations are performed. Model groups are created by dividing the population of establishments into sub-populations within which good regression fits can be achieved. This partition into model groups is based on combinations of NAICS industries.¹⁹ There are currently 75 model groups used.

The final estimates of SEPH use a combination of both data sources, the administrative and the BPS. Linear regression models are used to predict hours and summarized earnings from the sample of BPS respondents. Total employees and total payrolls for the month are the independent variables, while total hours and summarised earnings are the dependent variables. The regression coefficients estimated at the model group level are then applied to each record on the administrative source to mass impute hours and summarized earnings. Other variables (such as hourly employees or salaried employees) are obtained by multiplying employment, hours or summarized earnings by a ratio (or a function of ratios) estimated from the BPS sample.

The major difficulty with the use of administrative data for statistical purposes is that the raw information generally needs to be transformed in order to match the statistical concepts to be measured. In the case of the payroll deduction records, the main transformation is the conversion from reporting periods to reference periods. The estimates produced by SEPH should correspond to reference periods. Reporting requirements for the administrative source however varies according to the size of the business and the number of pay periods per year. The administrative data can represent any pay period (weekly, bi-weekly, etc.) within a reference month, or even across months that include the reference month. For instance, 13% of records correspond to weekly reporting periods, 20% are bi-weekly, 65% are monthly, and the remaining 2% have other reporting periods. Thus, 35% of the raw administrative data need to be transformed into (calendar) monthly data. Extensive sets of edit rules and transformations were developed to process the data. This resulted in a complete survey processing system that includes outlier detection, editing, imputation, and re-weighting.

The redesign of SEPH was so extensive and complex that it had to be staged in three phases. One downside to this multi-phased approach was that it affected the historical continuity of some of the estimates produced, notably total employment. Each phase of the redesign had an impact on the level of employment covered by the survey. For example, in Phase I the level of employment rose by 300,000 employees, while in Phase II the impact was in the order of 200,000 more employees. However, when Phase III of the redesign was implemented, where large businesses were included in the administrative sample, it was decided to delay the incorporation of the new levels of employment and payrolls to allow

¹⁹ Prior to the conversion to NAICS, model groups were based on combinations of industries and regions, and sometimes size. Since it was found that industrial activity was more discriminatory than geography, all model groups were kept at the national level with the conversion to NAICS.

more time for analysis, without delaying the reduction in response burden provided by the more comprehensive use of the administrative file.²⁰

The new levels of employment from Phase III were incorporated at the time of the conversion of SEPH to NAICS in January 2001. Hence, the level of employment for the year 2000 increased by about 285,000 employees. This new level of employment also reflects to a lesser extent slight modifications to the industrial coverage brought about by the conversion to NAICS, as well as improvements to the survey processes. As part of the conversion to NAICS the SEPH series were revised back to January 1991. This revision reduced many of the adverse impacts of the multi-staged redesign but could not remove all of them.

Over the last decade SEPH has gone through extensive changes designed to reduce respondent burden and improve its methodology. From a monthly sample survey of 70,000 establishments, SEPH gradually migrated to an administrative based survey supplemented with an 11,000 monthly sample survey of establishments, the Business Payrolls Survey. Since 1998, administrative data have contributed in significantly improving the efficiency of the survey, and major savings are made by using this wealth of information.

Use of Tax Data in the Monthly Restaurants, Caterers and Taverns Survey²¹

The Monthly Restaurants, Caterers and Taverns Survey (MRCTS) is a sample survey that collects sales and receipts and the number of locations from businesses in the Food Services and Drinking Places industry (NAICS 722). The principal activity of establishments in this industry is serving food and beverages. There are typically 56,000 such business locations in Canada. For the purposes of the survey, establishments are classified into five kinds of business: full service restaurants, limited service restaurants, food service contractors, social and mobile caterers, and drinking places.²²

The MRCTS was last redesigned in 1995, and last restratified in 1999. The MRCTS population is stratified by industry, region, and size. The frame for the survey is based on Statistics Canada's Business Register. Both the frame and sample are updated monthly, and are cumulative from month to month. The 'dead' units that do not appear in the monthly updates are kept in the frame to avoid biasing the estimates. There is no rotation of the respondents. Weights are computed every month within each stratum as the population count divided by the number of sample units (including dead units). This is equivalent to post-stratified estimation based on counts. The sample size is about 3,000.

As part of the tax data strategic streamlining initiative, Statistics Canada has started to review the MRCTS with a view toward reducing the response burden while strengthening the estimates produced in terms of population coverage and increased detail, even though its sample size is small by comparison with other surveys. The main tool at our disposal to achieve these goals is the Goods and Services Tax (GST) information available from Canada Customs and Revenue Agency. The object of this review is to use these administrative data as auxiliary information for deriving the estimates of sales for the industries covered by the survey. In effect, the purpose of this review is to use the GST information as a replacement for the survey data without redesigning the survey. The MRCTS will be one of the first monthly business surveys to use the GST information as auxiliary data. Hence, the experience gained with this survey will form a stepping-stone for broadening the use of the GST data to other surveys.

The Goods and Services Tax (GST) is a value-added tax that was introduced in 1991. It is remitted to Canada Customs and Revenue Agency (CCRA) from which it compiles administrative files that include the amounts of tax and of total sales reported by each business. Clearly, this information can be used to

²⁰ Since May 1998 and until December 2000, employment and payrolls growth rates were calculated from the administrative file and applied to the levels published in April 1998. This was to compensate the change in levels from Phase II to Phase III.

²¹ Much of this material is an abbreviated version of Hidiroglou, Dubreuil & Crowe (2003).

²² See Statistics Canada (2003) and www.statcan.ca/english/sdds/2419.htm.

great advantage by Statistics Canada to reduce costs and the burden of response while improving the detail and the quality of the estimates produced.

However, the raw information in these files must be processed to be statistically useful. One required treatment is the conversion of the information from reporting periods to calendar months. The files provided by CCRA to Statistics Canada are in the form of transactions representing taxable periods that may not exactly coincide with calendar periods. The reporting frequency is a function of the size of the business. Businesses report on a monthly, quarterly, or on an annual basis.²³ The reporting periods may also differ by establishment within an enterprise, and the reporting period of a business may change over time. About two-thirds of the MRCTS population represent 60% of the transactions reported on a quarterly basis.

Additionally, GST transactions are not all provided at once by CCRA to Statistics Canada for a given reference month. Although all transactions are eventually received from CCRA, a significant number need to be imputed the first time the data for a given reference month are processed. Tax Data Division at Statistics Canada performs this processing. It includes editing (i.e.: range edits, outliers), imputation (late returns, inconsistent values, critical outliers, partial and total non-response) as well as calendarization to account for the various filing frequencies (including extrapolation). Calendarization transforms all the GST data, including non-monthly data (quarterly, annual, 13 period), to correspond to the monthly period of business surveys.²⁴

Since the GST data for the current reference month are heavily imputed, we would run the risk of significant revisions if they were used to replace the survey data, or, but at a lesser extent, as auxiliary information. The following month however the information is updated by CCRA with the new data that they received. This means that for months prior to the reference month, much less imputation is required so that there is sufficient GST data for the previous reference month. Consequently, it is reasonable to use the calendarized GST data for the month prior to the survey reference month as auxiliary data for the current survey occasion.

The GST data will not be used at the present for large and complex enterprises. These enterprises will continue to be included in the survey, since disaggregating calendarized GST data from complex enterprises to the establishment level on a monthly basis can only be done on the basis of strong assumptions that may not hold over time. Modelling of the calendarized GST data will only be done for simple establishments.

The calendarized GST data may not exactly correspond to the data obtained by direct surveying. It is therefore necessary to adjust (model) them to correspond to survey data. In our case, GST should be more or less equal to a fraction of sales obtained from direct surveying, implying that the ratio estimator is a reasonable choice.²⁵ However, the estimator will be calculated over combined strata within the region by industry cells. Combining strata will contribute in increasing the efficiency of the ratio estimator by increasing the correlation between the survey data and the GST data.

One issue with the GST data is that they have a different coverage than the Business Register (BR), the frame used for the MRTCS. Establishments on the BR may be absent from the GST file because the BR is maintained from other sources of information, including information gathered by Statistics Canada's analysts. This is to be expected for units recently born and for dead units in the sample. Because of this, the MRCTS data cannot be entirely substituted with GST data. To capture the sales of those establishments not included in the GST file, we will use the same estimator used in the current MRCTS but computed over combined strata.

²³ Every business in Canada with \$30,000 or more in annual sales is required to register for a GST account with CCRA. Businesses with taxable sales of over \$6 million are required to report monthly. Those with taxable sales between \$500,000 and \$6 million have to report at least quarterly. Businesses with less than \$500,000 can report annually.

²⁴ See Quenneville, Cholette & Hidiroglou (2003).

²⁵ Since its inception the GST rate has been 7% for almost all goods and services sold by the industries covered by the MRCTS.

A related issue in using the GST data as auxiliary information for the MRCTS concerns the nature of the industries surveyed. This population is very dynamic, with large number of births and deaths between reference periods. Keeping track of all these changes is a crucial activity since the efficiency of the MRCTS design is largely linked to the quality of its frame information, particularly its accuracy and completeness. Dead units in the population are identified from a number of sources, and the survey declares in-sample unit deaths faster than the BR. As a result, the sample contains a larger proportion of identified dead units than the out-of-sample portion of the frame. Similarly, units appearing on the GST file can be declared as dead much later than their cessation of economic activities. The data for inactive GST records are imputed by a non-zero value for a set number of occasions as it is assumed that these units are still active.²⁶ As a result, if the corresponding in-sample unit is no longer active (a zero value), it is associated with a non-zero GST value which lowers the correlation between the GST information and the survey data. Dead units will nonetheless be included in the estimation, as is currently the case with the current survey, since doing so will stabilize the estimates.

Total sales will thus be estimated using a mixture of the combined post-stratified ratio estimator and the combined post-stratified count estimator. The combined post-stratified count estimator is used for sample subsets that cannot use the GST data, whereas the combined post-stratified ratio estimator is used for sample subsets that can use the GST data. The previous month GST data will be used as auxiliary data for the latter case as the number of non-imputed GST transactions is large enough to consider the resulting GST calendarized data as a reliable data source.

Studies on the use of the GST data in the context of the MRCTS are still in progress. Parallel runs of the current and modified surveys are expected to take place in the fall of 2003. The MRCTS augmented with the GST data is expected to be in operation sometime in 2004. Through the use of the Goods and Services Tax data, the response burden is anticipated to decrease by 25% to 35% for small establishments.

Redesign of the Monthly Wholesale and Retail Trade Survey²⁷

The Monthly Wholesale and Retail Trade Survey (MWRTS) produces estimates of sales and inventories for a large fraction of the Wholesale Trade (NAICS 41) and Retail Trade sectors (NAICS 44-45) by region and trade groups, i.e. special aggregations of industry classes. Together, the retail and wholesale sectors account for about 12% of Canada's GDP. The MWRTS is actually two distinct surveys, one for wholesale trade, and the other for retail trade. However, because of the similarities between these two sectors, their sample designs are similar and both surveys are designed (and redesigned) together.

There is a large population of establishments in wholesale and retail: around 106,000 for wholesale and 209,000 for retail. Both populations are highly skewed with few large businesses and many small ones. These populations are dynamic, with a large number of births, deaths and structural changes between reference periods. For example, in the retail sector, approximately 2,000 businesses are birthed each month and about the same number are declared out-of-business. Also, recent births are particularly unstable: approximately 20% of births in the wholesale and retail sector die during their first year of operation. Keeping track of all the changes in both populations is a crucial activity since the efficiency of the MWRTS sample design is largely linked to the quality of its frame information. However, the difficulties in tracking these dynamic populations must be taken into account when designing the survey.

The last redesign of the MWRTS occurred in 1988, and the survey was last restratified in 1997.²⁸ In 2000, the survey entered into a multi-year redesign process that will be staged in two phases. The purpose of this redesign is to convert from the 1980 Standard Industrial Classification to the 1997 North American Industry Classification System, and to use tax data information to produce a more efficient survey in

²⁶ Past this set number of occasions all previously imputed data are set to zero and the unit is assumed to be dead.

²⁷ Much of this material is an updated and abbreviated version of Bérard (2001). See also Ferland & Fortier (2003a,b).

²⁸ For the Wholesale Trade Survey, see Statistics Canada (2000a) and www.statcan.ca/english/sdds/2401.htm. For the Retail Trade Survey, see Statistics Canada (2000b) and www.statcan.ca/english/sdds/2406.htm.

terms of both costs and response burden. There is also the need to harmonise its concepts with the Annual Wholesale Trade Survey and the Annual Retail Trade Survey that have been recently redesigned.²⁹

Phase I of the redesign is in progress, and consists in the conversion to NAICS, the harmonisation of concepts, and the use of the Goods and Services Tax (GST) data, described in the previous section, to improve the stratification, particularly as regards the exclusion thresholds, so as to obtain a more efficient sample design that will reduce costs and the response burden. In Phase II, the GST data should more fully be used for sample replacement to estimate the relevant information for wholesalers and retailers in a fashion similar to that described in the previous section on the Monthly Restaurants, Caterers and Taverns Survey, where the tax replacement strategy is being applied to the survey design at it stands. In what follows only Phase I of the redesign of the MWRTS will be described.

As is the case for the current SIC-based MWRTS, the redesigned survey will not completely cover the Wholesale Trade and Retail Trade sectors. The main exclusions in wholesale are Oilseed and Grain Wholesalers (NAICS 41112), Petroleum Product Wholesalers (NAICS 412) and Wholesale Agents and Brokers (NAICS 419). The establishments excluded from retail are Non-Store Retailers (NAICS 454). As is the case for the current SIC-based survey, the redesigned NAICS-based MWRTS will publish estimates by trade groups, i.e. special aggregations of NAICS classes. There will be more of these however: 15 in wholesale and 19 in retail. Unlike the 1988 design, however, the survey population of the 2003 design will include both employers and non-employers. The 1988 design survey population only included employers, and an adjustment was performed to the estimates for retail to take into account the non-employers. Note that employers account for about 95% of the sales of both sectors.

Stratification by trade groups, regions, and establishments' size with simple random sample selection in each stratum will continue to be used for the new survey. However, many changes are brought to different aspects of the sampling design. One such change is that the sampling unit is no longer the company, but is defined as the cluster of establishments of the same enterprise that operate in the same industry group and same region.

A major change to the surveyed population will be the exclusion of the smallest establishments from sampling. This will be accomplished using exclusion thresholds applied at the industry-by-region cell.³⁰ The exclusion thresholds were designed such that the Goods and Services Tax (GST) data could be used to provide the estimates for these strata. The units excluded should account for less than 5% of the total sales for the cell. With the exclusions, the retail population is reduced by about 41%, from 209,000 to 124,000 establishments. The reduction in the wholesale population is even more drastic, at about 65%, decreasing from 106,000 to 37,000 establishments. Estimates for the non-surveyed portion will be produced using ratio estimation applied to the GST data.

Another change to the design is the use of a new measure of an establishment's size, developed specifically for the MWRTS. The measure is created using a combination of independent survey data and three administrative sources: the modelled Gross Business Income (GBI) on Statistics Canada's Business Register, the GST sales and the revenue declared on the Corporation Income Tax Return (T2). The independent survey data consist of the annual sales available from respondents to the current MWRTS (1988 design) and the annual wholesale and retail trade surveys. For respondents to the new MWRTS who also responded to one of these two surveys, the size measure will be set to the value of annual sales most recently reported, as this is deemed to be the better indicator of size. For the other respondents, the size measure will be set equal to the largest of the GBI, the GST sales and T2 revenue. For employers, a 10% reduction in the misclassification rate is observed when the size measure relies on the three administrative variables instead of the GBI alone. This new size measure is particularly efficient in identifying large businesses that should be classified to completely enumerated strata.

²⁹ See Parent & Simard (2000).

³⁰ See Statistics Canada (1998, 1999a).

Despite the added number of domains in the redesigned MWRTS and the precision required of the estimates, the sample size, and consequently the number of questionnaires, decreases substantially. For wholesale, the number of questionnaires to be sent will decrease from 8,200 to 5,600 (-32%). For retail, the number of questionnaires will decrease from 16,900 to 10,300 (-39%). These reductions are largely due to the application of the exclusion thresholds. The exclusion of a large portion of the target population from the survey population contributes to a more efficient sample design while reducing the burden of response for the smallest businesses. In an effort to reduce even further the response burden, sample rotation was originally planned for the new design. However, the various rotation schemes studied caused unacceptable fluctuations in the estimated monthly variations. Hence, rotation will not be implemented.

The five-month parallel run for the redesigned MWRTS is projected to start with the October 2003 reference month. The NAICS-based estimates from the survey will be published starting with reference month March 2004. At the same time, past SIC-based estimates converted to the new NAICS trade groups will be made available, from January 1991 for the Retail Trade sector, and from January 1993 for the Wholesale Trade Sector.

References

- Barnabé, R. (2003) *Seeing the Whole Elephant: A Proposed Experiment on Measuring the Activities of Multinational Enterprises*. Paper presented at the 2003 Conference of European Statisticians, June 2003. Statistics Canada. Available at www.unece.org/stats/documents/ces/2003/13.e.pdf.
- Baxter, W. (2003) *Statistics Canada's Business Respondent Relations*. Forthcoming. Statistics Canada.
- Bérard, H. (2001) The Redesign of the Monthly Wholesale and Retail Trade Survey of Statistics Canada. In *Proceedings of the Survey Methods Section*. Statistical Society of Canada.
- Bissett, P.D. (1999) *Use of Tax Data in the Production of Provincial Economic Statistics*. Paper presented at the 1999 Federal Committee on Statistical Methodology Conference, November 1999. Statistics Canada. Available at www.fcs.gov/99papers/bissett.html.
- Essoltani, A., Zorzi, N. (2003) *Building the Secure Internet Response Site (SIRS) and Its Supporting Systems*. Presentation at the 2003 Statistics Canada Information Technology Conference. April 2003.
- Ferland, M., Fortier, S. (2003a) *Retail – April 2003 Scenario*. Internal Statistics Canada document.
- Ferland, M., Fortier, S. (2003b) *Wholesale – April 2003 Scenario*. Internal Statistics Canada document.
- Gaudreau, M., Hughes, J. (2000) The Challenge of Collecting Quality Data From Large Enterprises: Lessons Learned From the Key Provider Manager Approach. In *ICES-II: Proceedings of the Second International Conference on Establishment Surveys*. Alexandria, VA : American Statistical Association.
- Grondin, C., Lavallée, P. (2001) *Survey of Employment, Payroll and Hours: An Update*. Internal Statistics Canada document.
- Hidiroglou, M.A., Dubreuil, G., Crowe, S. (2003) *Using the Goods and Services Tax (GST) for Monthly Surveys: Application to the Monthly Restaurants, Caterers, and Taverns Survey*. Document presented at the Advisory Committee on Statistical Methods, April 28-29, 2003. Statistics Canada.
- Leduc, J. (2001) *SEPH estimates are now based on the North American Industrial Classification System (NAICS)*. March 2001. Statistics Canada. Available at www.statcan.ca/english/sdds/document/2612_D2_T9_V1_E.pdf.
- Lee, H., Croal, J. (1991). *Use of PD-7 data for estimation in the Survey of Employment, Payrolls and Hours*. Working Paper BSMD-91-009E, Methodology Branch. Statistics Canada.
- Mayda, J. (2002) Experience with Implementation of EDR into Existing Survey Programs. *Statistical Journal of the United Nations Economic Commission for Europe*, 19(3), 131-140. Available at www.unece.org/stats/documents/2002.02.edr.htm.
- Parent, M.-N., Simard, M. (2000) Sampling with a Unified Approach: The Case of the Unified Enterprise Survey.

- Quenneville, B., Cholette, P.A., Hidioglou, M.A. (2003) *Estimating Calendar Month Values from Data with Various Reporting Frequencies*. Document presented at the Advisory Committee on Statistical Methods, April 28-29, 2003. Statistics Canada.
- Rancourt, E., Hidioglou, M.A. (1998) Use of Administrative Records in the Canadian Survey of Employment, Payrolls and Hours. In *Proceedings of the Survey Methodology Section*, Statistical Society of Canada.
- Schiopu-Kratina, I., Srinath, K.P. (1991) Sample Rotation and Estimation in the Survey of Employment, Payrolls, and Hours. *Survey Methodology*, 17, 79-90.
- St-Martin, P. (1992) Application of the General Regression Technique to Improve Estimates in an Establishment Survey. *Proceedings of the 1992 Annual Research Conference*, US Bureau of the Census, 225-244.
- Statistics Canada (1996) *A Proposal for a Unified Enterprise Statistics Program*. December 1996. Internal Statistics Canada document.
- Statistics Canada (1998) *Report of the Task Group on Data Acquisition for Enterprises*. Royce, D., Maranda, F. (co-chairs). July 1998. Internal Statistics Canada document.
- Statistics Canada (1999a) *Exclusions Thresholds and Specific Sampling Practices for Business Surveys – Implementation Strategy*. May 1999. Internal Statistics Canada document.
- Statistics Canada (1999b) *Unified Enterprise Survey Information Package*. Cat. No. 68F0015XIE. Available at www.statcan.ca/english/sdds/document/UES_D1_T1_V1_E.pdf.
- Statistics Canada (2000a) *Wholesale Trade Survey*. *Statistical Data Documentation System Reference Number 2401*. Available at www.statcan.ca/english/sdds/document/2401_D1_T2_V1_B.pdf.
- Statistics Canada (2000b) *Retail Trade Survey*. *Statistical Data Documentation System Reference Number 2406*. Available at www.statcan.ca/english/sdds/document/2406_D1_T2_V1_B.pdf.
- Statistics Canada (2001) *Guide to the Survey of Employment, Payrolls and Hours*. Cat. No. 72-620. Available at www.statcan.ca/english/sdds/document/2612_D1_T1_V1_E.pdf.
- Statistics Canada (2002) *Performance Report for the Period Ending March 31, 2002*. Ottawa: Treasury Board Secretariat. Available at www.tbs-sct.gc.ca/rma/dpr/01-02/SC/SC0102dpr_e.asp.
- Statistics Canada (2003) *Restaurant, Caterer and Tavern Statistics*. Monthly. Cat. No. 63-011-XIE.
- Tourigny, J., Pursey, S., Whitridge, P. (2002) The Unified Enterprise Survey. Its Approach to Quality. In *Symposium 2001 – Achieving Data Quality in a Statistical Survey: A Methodological Perspective. Proceedings of Statistics Canada's XVIIIth International Symposium on Methodological Issues and Workshops*. Cat. No. 11-522-XIE. Statistics Canada. Available at www.statcan.ca/english/conferences/symposium2002/session1/s1a.pdf.