Reporting Response Rates when Survey and Administrative Data are Combined

Julie Trépanier, Claude Julien and John Kovar

Statistics Canada
Tunney's Pasture, R. H. Coats Building, Ottawa (Ontario), K1A 0T6, Canada
julie.trepanier@statcan.ca, claude.julien@statcan.ca, john.kovar@statcan.ca

Abstract

At Statistics Canada, annual and monthly business surveys are using administrative data at an ever increasing rate. An important set of administrative data is received from the Canada Revenue Agency as a result of its collection of income tax reports and the Goods and Services Tax reports. These data are not only used to build and maintain a central frame, such as Statistics Canada's Business Register, or to assist in the imputation of survey data, they are now used to completely or partially replace subpopulations that would have traditionally been surveyed. The primary aim of the increased use is to reduce response burden and survey costs. Relying on a strong correlation between the administrative data and the survey data, the survey data can be either replaced directly with administrative data or indirectly through the production of modelled values based on the relationship between the two sets of data. As such, more and more business survey estimates are being based on a combination of survey and administrative information.

The traditional data quality indicators reported by surveys have been the sampling variance, coverage error, response rate and imputation rate. Are these indicators still relevant and sufficient in a context where administrative data are used? In fact, it is not unusual for some surveys to produce estimates that are based largely on the administrative source and thus reporting virtually no sampling error while other errors may be gaining in prominence: imputation error, model error, mode effects, etc. In 2004, a Task Force on Quality Indicators was set up at Statistics Canada to look into these issues and to recommend a strategy on how to report data quality in the context where survey and administrative data are combined. One of the main accomplishments achieved by the Task Force is a proposal to modify Statistics Canada's Standards and Guidelines for Reporting of Nonresponse Rates. This proposal is the focus of this paper. The use of a pie chart is also proposed as a visual means of simultaneously showing the different data sources and the response/nonresponse rates for each source.

1. Background

1.1 Integration of Survey and Administrative Data in Business Surveys

Section 24 of the Statistics Act has given Statistics Canada a specific right of access to federal income tax records from the Canada Revenue Agency since 1971. The last Memorandum of Understanding that clarifies the roles and responsibilities and outlines the conditions and procedures for the release of income tax and GST information between the two agencies was signed in April 2003. A decade ago, tax data were mostly used to build and maintain Statistics Canada's business survey frame (the Business Register), to support editing and imputation of survey data and for data confrontation. Starting in 1997, the annual business surveys program was expanded as a result of the Project to Improve Provincial Economic Statistics (Beelen et al. 1997). To offset the increase in response burden and costs, tax data were to be used as much as possible as a replacement for survey data traditionally collected from respondents. In this context, two main approaches have been implemented in business surveys: 1) the "take-none" approach, also referred to as the Royce-Maranda (R-M) approach (Royce and Maranda 1998), that excludes smaller businesses from regular data collection; and 2) the Tax Replacement (TR) approach that eliminates data collection for a significant portion of the sample of simple businesses. Both approaches are described in the next paragraphs. They are used in both our annual survey program and our three main monthly business surveys: the Monthly Restaurants, Caterers and Taverns Survey (MRCTS), the Monthly Survey of Manufacturing (MSM) and the Monthly Wholesale and Retail Trade Survey (MWRTS). The annual survey program relies on the income tax data (Nadeau 2004) while the monthly surveys rely on the Goods and Services Tax data (GST) (Dubreuil et al. 2003, Yung et al. 2004).

1.2 The Royce-Maranda (R-M) Approach

The R-M approach was first implemented for the 1998 reference year in the annual survey program. From a predetermined set of size thresholds, a threshold is chosen independently within each relevant industry by geography combination of a given survey. The threshold is determined so that the businesses below the threshold represent at most *x*% of a relevant size measure, where *x* is often set to 5% (in monthly surveys) or 10% (in annual surveys). These businesses make up the non-surveyed (or "take-none") portion of a survey. The remaining portion is called the surveyed portion (see Figure 1). The excluded businesses can be accounted for in the final estimates of key variables using tax data information in two manners. First, the non-surveyed businesses can be "micro" processed at each survey occasion (e.g., values are assigned to each individual business based on tax information) and then aggregated to produce a non-surveyed portion estimate to be added to the surveyed portion estimate. Secondly, some surveys may compute only a macro adjustment that is applied to the surveyed portion estimate to account for the planned "undercoverage" of the non-surveyed portion. For example, in Statistics Canada's MWRTS, the R-M approach was implemented with *x* set to 5%, resulting in respectively 68% and 45% of the wholesale and retail sampling units excluded from being surveyed. MWRTS currently computes a macro adjustment for the non-surveyed portion based on tax data information used at the time the survey was initially stratified. This constant macro adjustment is applied to the surveyed portion estimate to reflect the non-surveyed portion in the estimate.

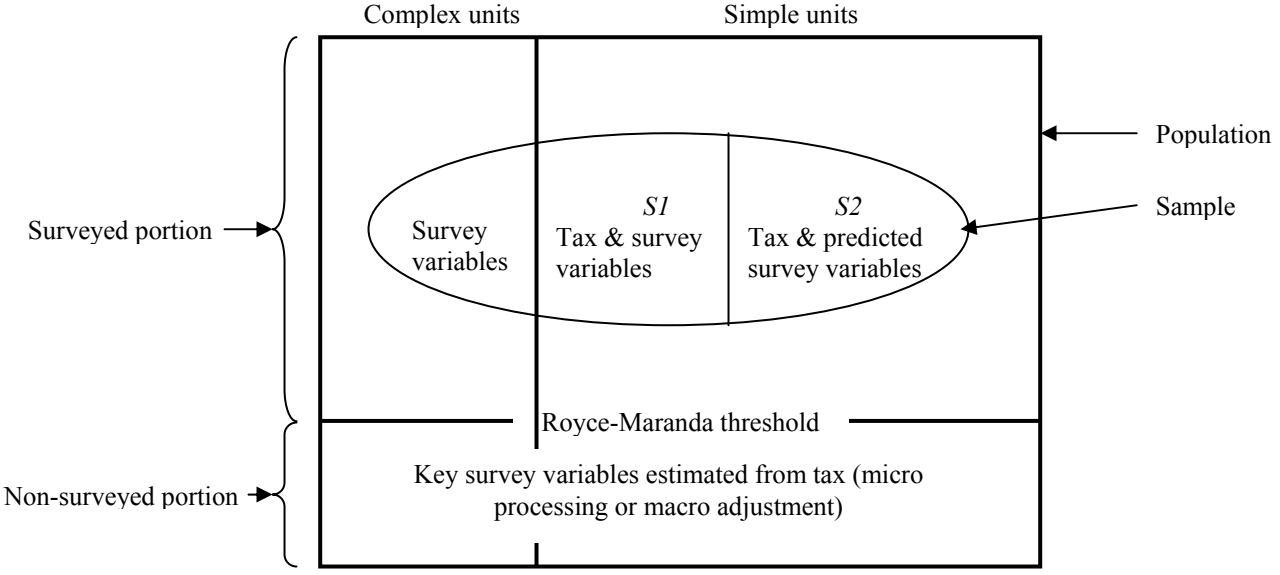
1.3 The Tax Replacement (TR) Approach

The TR approach took the use of tax data further starting with the 2002 reference year in the annual survey program and the 2004 reference year in the monthly surveys. In a given survey, units can be of two types of structure: simple or complex. In the case of simple units, the link between the unit and the level at which the tax data information is available can be easily established. Complex units are not as trivial and are not yet subjected to tax replacement. Hence, the TR approach only targets the sample of simple units in a given survey. The objective of the TR approach is to split the sample of simple units *S* into two parts, *S1* and *S2*. Survey data are still collected for the subsample *S1* and tax data are obtained for both *S1* and *S2*. No survey data are collected for *S2*; only tax data are available. Models are built using *S1* where tax variables are the explanatory variables and a survey variable is the dependent variable. The model parameters are then applied to the tax data variables in *S2* to predict the survey variables. The models are normally in the form of a simple linear regression model, a particular case being a direct substitution of the survey variable *Y* by the tax variable *X* (i.e., $\hat{Y}$=*X*). The size of *S2* varies from survey to survey but there is an attempt to put at least 50% of the sampled simple units in *S2*, thus reducing by half the response burden and collection costs of simple units. For example, 5,000 simple units are no longer collected by Statistics Canada in the three monthly surveys and instead their survey variables are predicted using the GST data. For the simple units, the TR translates into an estimator that is of the form of a one-phase imputed estimator $\hat{Y}_I$ of the total parameter *Y*, i.e.,

$$\hat{Y}_I = \sum_{i \in S} \pi_i^{-1} \tilde{y}_i$$

where π_i denotes the inclusion probability of unit *i* in the sample *S* and where $\tilde{y}_i = y_i$ (the survey value) if unit *i* is in *S1* and $\tilde{y}_i = y_i^*$ (the predicted survey value) if *i* is in *S2*.

Figure 1: Royce-Maranda and Tax Replacement approaches in a given business survey



1.4 Data Quality

Traditional data quality indicators reported by surveys are the sampling variance (or coefficient of variation, *cv*), response and nonresponse rates, imputation rate, and sometimes coverage rates. Most of the time, the sampling variance itself is approximated as all data are assumed reported when in fact some data are imputed. This was not too much of an issue when a survey such as the MWRTS had an imputation rate under 10% and the imputation methods, based on regression models, had proven to be efficient. However, the creation of a non-surveyed portion as represented in Figure 1 raised an issue: this portion is part of the survey (or observed) population although no sample is selected from it. There is no sampling error in the non-surveyed portion but there are definitely other sources of errors that the traditional indicators may not always illuminate. The latter is particularly true because Statistics Canada's current Standards and Guidelines for Reporting of Nonresponse Rates (Statistics Canada 2001) clearly state the following:

> The standards apply to censuses and sample surveys which are based on direct data collection from respondents. They do not apply to surveys based only on administrative records. However, for surveys based on direct data collection for some units and administrative records for other units, the standards apply to the portions of the survey based on direct data collection.

When only the R-M approach was used, most surveys complied with the Standards and Guidelines by producing response and nonresponse rates that referred to the surveyed portion only. This made some sense, especially when the non-surveyed portion was accounted for in the estimate through a macro adjustment to the surveyed portion estimate. This became difficult to do when the TR approach was implemented in surveys. Excluding *S2* (see Figure 1) from the scope of the response and nonresponse rates would diminish some of the main purposes of the response and nonresponse rates. For instance, it is generally understood that weighted response and nonresponse rates are used to indicate the proportion of an estimate that is contributed by respondents. This is hard to achieve when an estimate (based on both surveyed and tax units) and the weighted response rate (based on surveyed units) do not have the same reference base. Another difficulty was to decide on the status of the *S2* units and the non-surveyed portion units (in the situation where the latter are micro-processed). The substitution by tax data (direct or via a model) was considered by some as another mode of collecting the information, and by others as imputed data.

1.5 2004-2005 Statistics Canada's Task Force on Quality Indicators

The context explained in 1.4 led to the creation of a Task Force on Quality Indicators in 2004. In the last year, the Task Force accomplished the following: 1) definition of its framework; 2) proposal to modify Statistics Canada's Standards and Guidelines for Reporting of Nonresponse Rates; 3) proposal of visual means to better understand the integration of survey and tax data in a survey and the associated sources of errors; and 4) proposal on how to produce combined imputation rates when survey and tax data are used (Lavallée 2005).

The present paper[1] briefly describes (1) and mainly focuses on (2). The use of a pie chart is also presented as a visual means of simultaneously showing the different data sources and the response/nonresponse rates for each source.

The integration of survey and tax data has many more issues than just the definition of quality indicators. Since using more tax data (and perhaps in a more efficient way) is seen as one of the main priorities of Statistics Canada, a new project, the Data Integration Project (DIP), was launched in 2005. Future activities related to data quality indicators when both survey and administrative data are integrated would most likely fall under the DIP.

2. Task Force on Quality Indicators

2.1 Definition of its Framework

The Task Force on Quality Indicators first established a few principles that would be the basis for its work.

Focus on business surveys: Although administrative data are also used in social surveys, the developmental activities performed by the Task Force were to be done in the context of business surveys as described in section 1 of this paper. However, the results might be applicable to social surveys.

---

[1] A complete report will be prepared by the Statistics Canada's Task Force on Quality Indicators, i.e., John Kovar (initial chairperson), Claude Julien (co-chairperson), Julie Trépanier (co-chairperson), Hélène Bérard and Pierre Lavallée.

Focus on the "accuracy" dimension of quality: Statistics Canada's Quality Assurance Framework (Statistics Canada 2002) describes six dimensions of quality: accessibility, accuracy, coherence, interpretability, relevance and timeliness. Although the work of the Task Force may affect other dimensions of the quality of an end-product, developmental activities performed by the Task Force are meant to provide better indicators of the accuracy of the estimates produced by a survey that combines survey and tax data: both sampling errors as well as nonsampling errors such as coverage, nonresponse, reporting, model, etc.

Quality indicators from the data producers' perspective: Quality indicators are produced for three groups of stakeholders: the users, the managers and the producers. The producers (members of the survey team) are normally the ones who require the greatest number of quality indicators and the most detailed ones. A subset of these indicators is provided to the managers (usually those who control the budgets). A different subset of the indicators is made available to data users to help them establish whether the data are appropriate for their purposes. The Task Force made the decision to go with the wider perspective (the producers' one) since a subset of the indicators can be used for the managers and users.

Under that framework, the Task Force decided to look in particular at the issue of response and nonresponse rates. It aimed at developing an approach based on Statistics Canada's Standards and Guidelines for Reporting of Nonresponse Rates. The resulting rates should indicate in particular: 1) the proportions of an estimate that come from different sources; 2) the proportions within each source that are based on reported and nonreported data. These two aspects could then be expressed as a pie chart. This is described in the following paragraphs. The next steps for a given survey would be to identify the different errors associated with each source and to try to estimate the main sources of errors using appropriate techniques.

2.1 Proposal to Modify the Standards and Guidelines for Reporting of Nonresponse Rates

The Guidelines for Measuring Statistical Quality produced by the Office for National Statistics (Office for National Statistics 2004) provide an interesting comparison between quality measures and quality indicators:

> Quality of data can rarely be explicitly « measured » (…) Quality indicators usually consist of information which is a by-product of the statistical process. They do not measure quality directly, but can provide enough information to make inferences about the quality.

In that sense, response and nonresponse rates are quality indicators and not quality measures. A nonresponse rate never replaces attempts that are made to measure nonresponse bias or to estimate variance in the presence of nonresponse and imputation. However, they are useful and easy to understand when properly defined. They may raise quality concerns about the data and convince the data producers to improve the data collection procedures, to conduct a nonresponse study or to measure variance taking into account nonresponse and imputation.

The scope of Statistics Canada's current Standard and Guidelines for Reporting of Nonresponse Rates does not include surveys or portions of a survey that are based on administrative records. In what follows, we first provide an overview of the Standards and Guidelines. Secondly, we describe how they could be modified to the context where both survey and tax data are used in a given survey. It is important to note that before the proposal can become official Standards and Guidelines it must go through a number of approval steps at Statistics Canada.

2.1.1 Standards and Guidelines for Reporting of Nonresponse Rates. The Standards and Guidelines were developed to standardize the ways response and nonresponse rates were calculated across surveys. The intent was to allow comparison between surveys, and analysis of trends in data collection and respondent behaviour over time within as well as across surveys.

Response and nonresponse rates are calculated at two phases: the collection phase and the estimation phase. The primary purpose of the Standards and Guidelines is to calculate unit rates although they can be also applied to item rates (i.e., questionnaire item rates). At the collection phase, the choice of the unit of reference in business surveys is normally the collection unit as defined by Statistics Canada's Business Register. The same unit can be used at the estimation phase although the "statistical" unit may be chosen as the unit of reference. Because business survey populations are often highly skewed, it is recommended to calculate both unweighted and weighted rates. For the weighted rates, one could use the sampling weight, but it is usually preferable to combine it with an economic weight (i.e., a key size variable available for all sampled units).

The framework for calculating the rates at the collection phase is based on the classification of units in the survey into a nested hierarchy of categories. The main current classification is graphically presented in Appendix A of this paper, with boxes delimited by a solid line. The main categories in the Standards and Guidelines are defined as follows:

Total Units:	All units included in the census or sample survey.
Resolved Units:	Units whose status has been resolved by the end of the period of survey data gathering, as either belonging or not belonging to the target universe for the survey.
Unresolved Units:	Units whose status has not been resolved by the end of the period of survey data gathering.
In-scope Units:	Resolved units determined to belong to the target universe for the survey.
Out-of-scope Units:	Resolved units determined to not belong to the target universe for the survey. This category can be further broken down (not shown in the Appendices A and B) into non-existent units ("deaths"), temporarily out-of-scope units (e.g., seasonally closed businesses) and permanently out-of-scope units (e.g., not in the industry or geography targeted by the survey).
Responding Units:	In-scope units which, at the data collection phase, are deemed to have responded by virtue of having provided usable information. Consequently, it is necessary at this phase to define a survey-specific threshold for "usable information" in terms of the level to which incomplete questionnaires have to be filled out before a unit is classified as "responding".
Nonresponding Units:	In-scope units that are either nonrespondents or those that provide information below the "usable" threshold. This category can be further broken down into "no contacts", refusals and residual nonresponding units.

The key rate, i.e., the response rate for data collection, is calculated as:

$$\text{Response rate for data collection} = \frac{\text{Responding Units}}{\text{In-scope Units + Unresolved Units}} = \frac{\text{Box C.1.1.1}}{\text{Box C.1.1 + Box C.2}}$$

It is important to note that resolved out-of-scope units are excluded from both the numerator and denominator. Including resolved out-of-scope units in the calculation of the response rate for data collection would artificially increase the response rate, in particular if there are a large number of out-of-scope units in the survey.

Here and everywhere else in this paper, the equivalent nonresponse rate is *(1 – response rate)*. Also, the rates can be weighted by using the sampling weight in order to illuminate possible effects or dependencies on the survey design. Weighting by a size variable (economic weight) is rarely useful at the collection stage (but it is at the estimation stage), though it could be used during the collection operation along with a suitable score function in order to target more important units.

At the estimation phase, the Standards and Guidelines define additional categories (see Appendix B – boxes with solid lines). They are the following:

Estimated In-scope Units:	Unresolved units that are estimated to be in-scope. This can be based on the latest information available from the survey, other surveys, frames, administrative files, etc. The basis for defining if a unit is estimated to be in-scope (or out-of-scope) should be the same as the one used when estimates are produced by the survey.
Estimated Out-of-scope Units:	Unresolved units that are not estimated to be in-scope.
Unusable:	Units that were considered responding at the data collection phase, but for estimation purposes, were discovered to be unusable because the level of completion of the questionnaire and/or the level of errors noted at the processing stage made these units fall below the survey-specific threshold.
Usable:	Units that were considered responding at the data collection phase whose level of information is still deemed usable according to the survey-specific threshold.

The key rate, i.e., the response rate for estimation, is calculated as:

$$\text{Response rate for estimation} = \frac{\text{Responding Units} - \text{Unusable}}{\text{In-scope Units} + \text{Estimated In-scope Units}} = \frac{\text{Box C.1.1.1} - \text{Box C.1.1.1.2}}{\text{Box C.1.1} + \text{Box C.2.1}}$$

When weighted, the rates should be using both the sampling weight and an economic weight. Weighting by the sampling weight alone is not as useful in business surveys but is (more) appropriate for social surveys.

2.1.2 Proposed Additions to the Standards and Guidelines. The Task Force chose to consider the use of administrative data as another mode of collecting the data. However, it is important to realize that the statistical organisation has normally little control on how the administrative data are being collected by the administrative program. This fact led the Task Force to expand the categories proposed in the Standards and Guidelines while adapting them to the context of administrative records. The first step is to clarify the definition of the "Total Units" category.

Total Units: Composed of any unit included in the census or sample survey that is meant by design to be observed either by direct collection (Box C in Appendix A) or by data extraction from an administrative file (Box A in Appendix A). In a situation where both administrative data and data collected directly are to be obtained for a portion of the sample (e.g., *S1* in Figure 1, for the purposes of calculating model parameters), one needs to classify the units in Box A or Box C according to the origin of the data used for these units at estimation. This means in our example that *S1* units fall in Box C (because the data collected directly from the respondents is used at estimation for units in *S1*).

The word "observed" in the previous definition is a key one. Let us take the example of the non-surveyed portion as described in Figure 1. As mentioned in section 1.2, this portion often represents 5 to 10% of a survey-specific size measure. For a given survey (e.g., a multivariate survey where variables are more or less correlated), it may be more appropriate to obtain tax information for all units included in this non-surveyed portion. In this situation, the units in the non-surveyed portion are meant to be observed individually. The tax data may or may not be obtained. The survey then needs to decide which units are deemed in-scope and out-of-scope. The tax data will be edited and imputed and ultimately an estimate will be produced for that non-surveyed portion. At the opposite end of the spectrum, the non-surveyed portion could be seen by another survey as a planned undercoverage in the survey. The survey does not intend to observe each individual unit in the non-surveyed portion but instead, based on tax data for example, the survey plans to calculate a macro adjustment that will be applied to the surveyed portion estimates to account for this planned undercoverage. In the first scenario, the units in the non-surveyed portion are included in the "Total Units" category. In the second scenario, they are not.

At Statistics Canada, most business surveys use the Business Register to extract their frame. Frames are rarely perfect and the Business Register is no exception. It may contain inactive businesses or businesses that are misclassified with respect to industry and/or geography codes. If a particular business survey uses the R-M and TR approaches, they will be applied to this imperfect frame. Once the tax file is ready for use by business surveys, the extraction of all the records for the units selected by design to be "extracted from an administrative file" may not always be successful. It appeared to the Task Force that the categories "Resolved Units" and "Unresolved Units" that are used in direct collection were not appropriate to use for administrative records. The main reason is that "Resolved" means that one has confirmed that the unit is in-scope or out-of-scope. One could think that finding a record on a tax file means the record is in-scope. To a certain extent, it does show the unit is active but one can rarely determine with certainty (at least in the business survey world) that the unit still belongs to the industry and geography targeted by the survey. Our experience in Canada is that industry codes that may be present on tax files are sometimes either of poor quality or outdated. Along the same line, one could think that the fact that a record is not found in the tax file indicates the unit does not exist anymore. This is not always true as many reasons can explain the "absence" of a unit in a tax file such as the unit did not or was not yet required to provide its tax information, or problems were encountered in the record linkage process, etc.. As a result, assumptions about whether a unit that exists or not on the tax file is in-scope or out-of-scope needs to be made. It is our recommendation to address this in the estimation phase categories. Consequently, only two categories are created at the data collection phase under Box A:

Extracted Units: Units for which the intention (i.e., by design) was to extract the information from an administrative file and for which the extraction was successful.

Unextracted Units: Units for which the intention was to extract the information from an administrative file and for which the extraction was not successful.

Consequently, two key rates are calculated for the data collection phase. The first one still describes the quality of the collection effort that is under the control of the statistical agency. The second one describes the success in extracting the required units from the administrative file.

$$\text{Response rate for direct data collection} = \frac{\text{Responding Units}}{\text{In-scope Units + Unresolved Units}} = \frac{\text{Box C.1.1.1}}{\text{Box C.1.1 + Box C.2}}$$

$$\text{Extraction rate from the administrative file} = \frac{\text{Extracted Units}}{\text{Extracted Units + Unextracted Units}} = \frac{\text{Box A.1}}{\text{Box A.1 + Box A.2}}$$

At the estimation stage, it was already mentioned that for both the "extracted" and "unextracted" categories, one needs to estimate the number of units deemed to be in-scope and out-of-scope (Boxes A.1.1, A.1.2, A.2.1 and A.2.2 in Appendix B). Similarly to the data obtained through direct collection, administrative data that are "extracted" and deemed "in-scope" are further divided into four categories (see Appendix B).

Reporting Units:	Units estimated in-scope for which sufficient reported tax information is obtained. Consequently, it is necessary at this phase to define a survey-specific threshold related to the concept of "sufficient reported tax information". As an example, the monthly surveys that use the GST information require that the GST revenues be reported on the GST file to declare a unit as "reporting". This variable is not always present on the file since this variable is not a key one for the administrative program (the GST collected, i.e., the amount of GST tax remitted to the Canada Revenue Agency is its key variable).
Nonreporting Units:	Units estimated to be in-scope but are not considered reporting units.
Unusable:	Units that were first considered reporting but once the processing was performed were found to be incoherent or not plausible. To use the same example as before, the GST revenues may be present for a given unit, but once processed, may be found to be inconsistent with the GST collected or with the historical series of GST revenues of this unit.
Usable:	Units that were considered reporting and were still found usable after the processing of the data.

It appears relevant to the Task Force to produce rates at the estimation phase for the "Collected Directly" portion and the "Extracted from an Administrative Data File" portion separately as well as combined rates. The suggested rates are:

$$\text{Response rate for estimation (Direct collection)} = \frac{\text{Responding Units – Unusable}}{\text{In-scope Units + Estimated In-scope Units}} = \frac{\text{Box C.1.1.1 – Box C.1.1.1.2}}{\text{Box C.1.1 + Box C.2.1}}$$

$$\text{Response rate for estimation (Extracted from an Administrative file)} = \frac{\text{Reporting Units – Unusable}}{\text{Extracted Estimated In-scope Units + Unextracted Estimated In-scope Units}} = \frac{\text{Box A.1.1.1 – Box A.1.1.1.2}}{\text{Box A.1.1 + Box A.2.1}}$$

$$\text{Response rate for estimation (combined)} = \frac{\text{(Responding Units – Unusable) + (Reporting Units – Unusable)}}{\text{(In-scope Units + Estimated In-scope Units) + (Extracted In-scope Units + Unextracted In-scope Units)}} = \frac{\text{(Box C.1.1.1 – Box C.1.1.1.2) + (Box A.1.1.1 – Box A.1.1.1.2)}}{\text{(Box C.1.1 + Box C.2.1) + (Box A.1.1 + Box A.2.1)}}$$

Lavallée (2005) also presents options to produce combined imputation rates when both administrative and survey data are used. In particular he proposes rates that are applicable when the classification of all units into responding, reporting, usable etc cannot be performed but response rates are known at a certain grouping level for either one of the sources.

In parallel to these "estimation" response rates, the Task Force found useful to define "source" rates, i.e., the proportion of in-scope and estimated in-scope units that comes from each source. When appropriately weighted, these rates indicate the proportion of the total estimate that comes from each source.

$$\text{“Collected directly” source rate} = \frac{\text{(In-scope Units + Estimated In-scope Units)}}{\text{(In-scope Units + Estimated In-scope Units) + (Extracted In-scope Units + Unextracted In-scope Units)}}$$

$$= \frac{\text{(Box C.1.1 + Box C.2.1)}}{\text{(Box C.1.1 + Box C.2.1) + (Box A.1.1 + Box A.2.1)}}$$

$$\text{“Extracted from Admin” source rate} = \frac{\text{(Extracted In-scope Units + Unextracted In-scope Units)}}{\text{(In-scope Units + Estimated In-scope Units) + (Extracted In-scope Units + Unextracted In-scope Units)}}$$

$$= \frac{\text{(Box A.1.1 + Box A.2.1)}}{\text{(Box C.1.1 + Box C.2.1) + (Box A.1.1 + Box A.2.1)}}$$

Using the sampling weight and, as an economic weight, a key survey variable in both the weighted estimation response rates and the source rates allows one to build a pie chart that provides a good overview of the composition of this key variable estimate. This is shown in 2.1.3 where the proposed additions to the Standards and Guidelines are applied to the Monthly Wholesale and Retail Trade Survey.

Before ending this section, it is important to distinguish the following: It is possible that a unit was intended to be collected directly, and this unit may end up being either: 1) unresolved but estimated to be in-scope; or 2) in-scope and responding but unusable. The end-result may be that this unit's contribution will be imputed by using tax information and methods that are very similar to the way the tax information is used for units that were intended to be extracted from that same tax file. On one hand, the unit contributes to the nonresponse rate; on the other hand, it contributes to the response rate. At first, this may seem inconsistent. However, one has to keep in mind that the unit that was to be collected directly and happened not to respond hides a nonresponse mechanism that is unknown in most cases with the potential of creating nonresponse bias. This is not the case for units intended to be extracted from the tax file. They were chosen by design by the survey statistician, under a well known selection scheme.

2.1.3 Application in the Monthly Wholesale and Retail Trade Survey. MWRTS is currently testing the TR approach using GST data in parallel with its regular production. The production side is still based on a sample in the surveyed portion that is totally intended to be collected directly from the respondents. On the test side for April 2005 reference month, the sample in the surveyed portion for Retail was made up of 11,579 collection units; 9,973 were intended to be collected directly, predicted sales based on March 2005 GST information (via model parameters calculated in *S1*) were calculated for the other 1606 units. Meanwhile for Wholesale, the sample in the surveyed portion was made up of 6,088 collection units; 5,209 were intended to be collected directly, 879 received predicted sales based on March 2005 GST information. The non-surveyed portion of MWRTS is considered to be non-observed (and not part of the "Total Units" category); a macro adjustment is made to the final estimate to account for this planned undercoverage. At the estimation phase, the following numbers were taken into account by categories on the test side. The equivalent weighted values (weighted by the main variable of the survey, the total sales, and the sampling weight) are provided as well. Please note again that these numbers were produced during a test for the April 2005 reference month and are unofficial estimates. Official estimates released by Statistics Canada can be found at www.statcan.ca.

Table 1A: April 2005 TR Test on the Monthly Retail Trade Survey: Counts and Weighted Values for the Purpose of Calculating Estimation Phase Response and Source Rates

Categories	Counts	Weighted value (in millions $)
C. Collected directly	9,973	27,315
C.1 Resolved Units	9,973	27,315
C.1.1 In-scope Units	6,178	27,315
C.1.1.1 Responding Units	5,195	24,122
C.1.1.1.1 Usable	5,186	24,023
C.1.1.1.2 Unusable	9	99
C.1.1.2. Nonresponding Units	983	3,193
C 1.2 Out-of-scope Units	3,795	0
C.2 Unresolved Units	0	0
C 2.1 Estimated In-scope Units	0	0

C 2.2 Estimated Out-of-scope Units	0	0
A. Extracted from an Administrative (GST) Data File	1,606	2,807
A.1 Extracted Units	1,596	2,800
A.1.1 Estimated In-scope Units	1,596	2,800
A.1.1.1 Reporting Units	1,387	2,513
A.1.1.1.1 Usable	1,194	2,283
A.1.1.1.2 Unusable	193	230
A.1.1.2 Nonreporting Units	209	287
A 1.2 Estimated Out-of-scope Units	0	0
A.2 Unextracted Units	10	8
A.2.1 Estimated In-scope Units	5	8
A.2.2 Estimated Out-of-scope Units	5	0

For Retail, the weighted response and source rates at the estimation phase are then:

$$\text{Response rate for estimation (Direct collection)} = \frac{\text{Box C.1.1.1} - \text{Box C.1.1.1.2}}{\text{Box C.1.1} + \text{Box C.2.1}} = \frac{24{,}122 - 99}{27{,}315 + 0} = 87.9\%$$

$$\text{Response rate for estimation (Extracted from Admin. File)} = \frac{\text{Box A.1.1.1} - \text{Box A.1.1.1.2}}{\text{Box A.1.1} + \text{Box A.2.1}} = \frac{2{,}513 - 230}{2{,}800 + 8} = 81.3\%$$

$$\text{Response rate for estimation (Combined)} = \frac{(24{,}122 - 99) + (2{,}513 - 230)}{(27{,}315 + 0) + (2{,}800 + 8)} = 87.3\%$$

$$\text{“Collected directly” source rate} = \frac{(\text{Box C.1.1} + \text{Box C.2.1})}{(\text{Box C.1.1} + \text{Box C.2.1}) + (\text{Box A.1.1} + \text{Box A.2.1})} = \frac{(27{,}315 + 0)}{(27{,}315 + 0) + (2{,}800 + 8)} = 90.7\%$$

$$\text{“Extracted from admin.” source rate} = \frac{(\text{Box A.1.1} + \text{Box A.2.1})}{(\text{Box C.1.1} + \text{Box C.2.1}) + (\text{Box A.1.1} + \text{Box A.2.1})} = \frac{(2{,}800 + 8)}{(27{,}315 + 0) + (2{,}800 + 8)} = 9.3\%$$

By taking the source rates and multiplying each of them by the weighted response rate associated to it, we can show the composition of the total estimate using a pie chart.

Figure 2A: April 2005 TR Test for the Monthly Retail Trade Survey
Composition of the total sales estimate

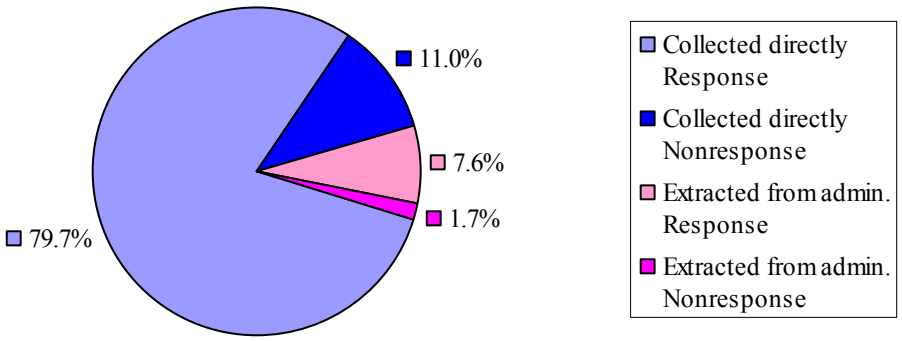


Table 1B: April 2005 TR Test on the Monthly Wholesale Trade Survey: Counts and Weighted Values for the Purpose of Calculating Estimation Phase Response and Source Rates

Categories	Counts	Weighted value (in millions $)
C. Collected directly	5,209	35,750
C.1 Resolved Units	5,209	35,750
C.1.1 In-scope Units	3,547	35,750
C.1.1.1 Responding Units	3,041	32,297
C.1.1.1.1 Usable	3,036	32,227
C.1.1.1.2 Unusable	5	70
C.1.1.2. Nonresponding Units	506	3,453
C 1.2 Out-of-scope Units	1,662	0
C.2 Unresolved Units	0	0
C 2.1 Estimated In-scope Units	0	0
C 2.2 Estimated Out-of-scope Units	0	0
A. Extracted from an Administrative (GST) Data File	879	3,575
A.1 Extracted Units	873	3,574
A.1.1 Estimated In-scope Units	873	3,574
A.1.1.1 Reporting Units	743	3,093
A.1.1.1.1 Usable	601	2,525
A.1.1.1.2 Unusable	142	569
A.1.1.2 Nonreporting Units	130	480
A 1.2 Estimated Out-of-scope Units	0	0
A.2 Unextracted Units	6	1
A.2.1 Estimated In-scope Units	4	1
A.2.2 Estimated Out-of-scope Units	2	0

For Wholesale, the weighted response and source rates at the estimation phase are then:

$$\text{Response rate for estimation (Direct collection)} = \frac{\text{Box C.1.1.1} - \text{Box C.1.1.1.2}}{\text{Box C.1.1} + \text{Box C.2.1}} = \frac{32{,}297 - 70}{35{,}750 + 0} = 90.1\%$$

$$\text{Response rate for estimation (Extracted from Admin. File)} = \frac{\text{Box A.1.1.1} - \text{Box A.1.1.1.2}}{\text{Box A.1.1} + \text{Box A.2.1}} = \frac{3{,}093 - 569}{3{,}574 + 1} = 70.6\%$$

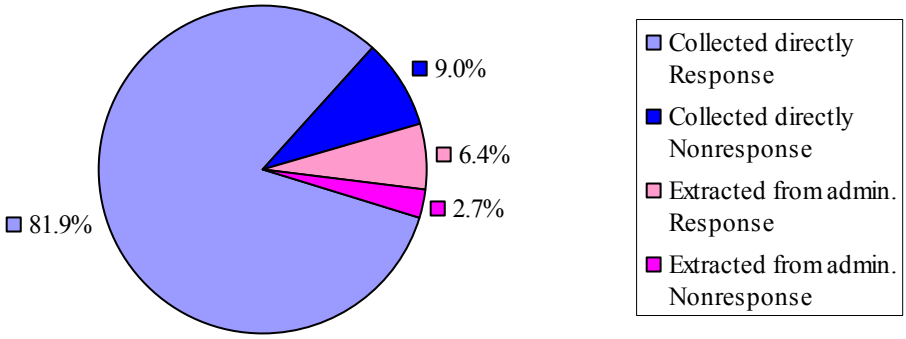
$$\text{Response rate for estimation (Combined)} = \frac{(32{,}297 - 70) + (3{,}093 - 569)}{(35{,}750 + 0) + (3{,}574 + 1)} = 88.4\%$$

$$\text{“Collected directly” source rate} = \frac{(\text{Box C.1.1} + \text{Box C.2.1})}{(\text{Box C.1.1} + \text{Box C.2.1}) + (\text{Box A.1.1} + \text{Box A.2.1})} = \frac{(35{,}750 + 0)}{(35{,}750 + 0) + (3{,}574 + 1)} = 90.9\%$$

$$\text{“Extracted from admin.” source rate} = \frac{(\text{Box A.1.1} + \text{Box A.2.1})}{(\text{Box C.1.1} + \text{Box C.2.1}) + (\text{Box A.1.1} + \text{Box A.2.1})} = \frac{(3{,}574 + 1)}{(35{,}750 + 0) + (3{,}574 + 1)} = 9.1\%$$

As in Retail, the composition of the Wholesale estimates can be shown in a pie chart.

Figure 2B: April 2005 TR Test for the Monthly Wholesale Trade Survey
Composition of the total sales estimate



3. Conclusion

Business surveys at Statistics Canada are making increasing use of tax data as a replacement to direct data collection, either by creating a non-surveyed portion that is totally estimated from tax data or by selecting a subsample of the sample of simple units to be modelled from tax data. This has triggered the need to review our way of reporting data quality, namely response and nonresponse rates. This paper has proposed a way to adapt these quality indicators. In particular, the rates computed at the estimation phase, when properly weighted, have the advantage of explaining the proportion of a key variable estimate that comes from each source ("Collected directly" and "Extracted from an administrative data file") and the proportion of each source that is based on responded/reported data. These rates can be represented visually by a pie chart. They are meant to be quality *indicators* and not quality *measures*. As such, one should use these rates as a starting point to identify the main sources of errors associated with each data source and to attempt to measure these errors.

As for Statistics Canada's Task Force on Quality Indicators, a report will be tabled later in 2005. The proposal presented here to modify the Standards and Guidelines for Reporting of Nonresponse Rates will be put to the test in other surveys besides the MWRTS, which was presented in this paper. If the test is successful, the resulting proposal would be presented to the Statistics Canada's Methods and Standards Committee for their advice and approval.

Finally, a new project entitled the Data Integration Project (DIP) (Trépanier 2005) was launched in April 2005. It is a three-year methodology project that aims at researching and developing efficient methods to push further the integration of subannual (like GST data) and annual (like the income tax data) administrative data as well as subannual and annual survey data for a given statistical program. Methods researched and developed must produce accurate, relevant, timely and coherent estimates within a statistical program. The resulting methods may or may not be an extension of the current R-M and TR approaches. Calibration techniques are an avenue being looked at.

4. Acknowledgements

The authors would like to thank Hélène Bérard and Pierre Lavallée who significantly contributed to the accomplishments made by the Task Force. Special thanks to Doug Batten who extracted and computed the necessary information for the MWRTS application. Finally, sincere thanks to the reviewers, Pierre Lavallée, Mark Majkowski and Don Royce who provided relevant comments on the initial draft of this paper.

5. References

Beelen, G., Hardy, F. and Royce, D (1997). "An overview of the Project to Improve Provincial Economic Statistics". Internal document, Statistics Canada.

Dubreuil, G., Hidiroglou, M. A. and Pierre, L. (2003). "Use of Administrative Data in Modeling of the Monthly Survey Data". *Proceedings of the Survey Methods Section,* Statistical Society of Canada, CD-Rom.

Lavallée, P. (2005). "Quality Indicators when Combining Survey Data and Administrative Data". *Proceedings of the XXII International Methodology Symposium,* Statistics Canada, CD-Rom (to appear).

Nadeau, C. (2004). "Challenges Associated with the Increased Use of Fiscal Data for the Unified Enterprise Survey." *Proceedings of the Section on Survey Research Methods,* American Statistical Association, CD-Rom.

Office for National Statistics (2004). *Guidelines for Measuring Statistical Quality: Version 1.0*. Office for National Statistics.

Royce, D. and Maranda, F. (1998). "Task Force on Business Data Acquisition", Internal Report, Statistics Canada.

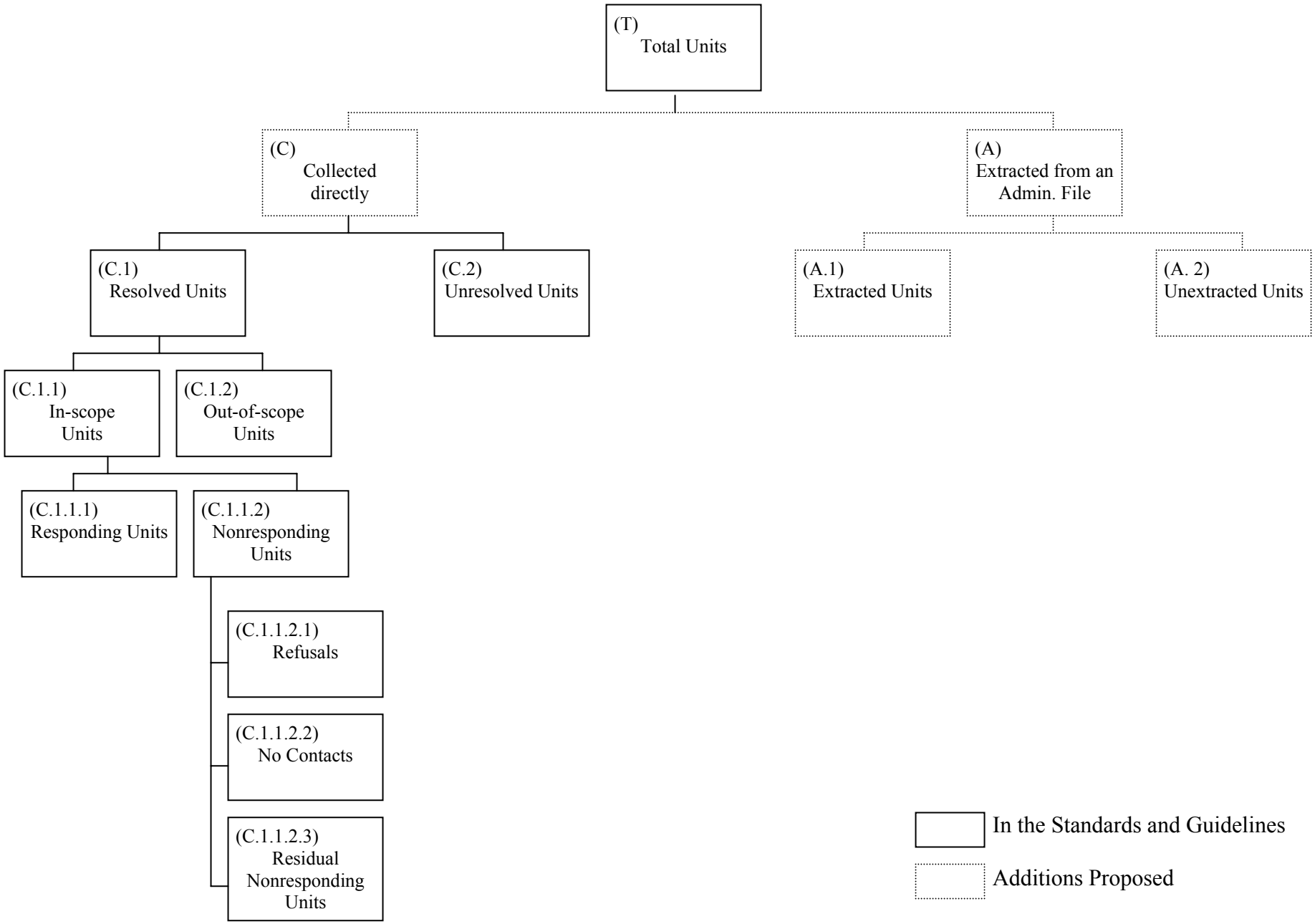
Statistics Canada (2001). "Standards and Guidelines for Reporting of Nonresponse Rates: Definitions, Framework and Detailed Guidelines". Internal document, Statistics Canada.

Statistics Canada (2002). *Statistics Canada's Quality Assurance Framework*. Statistics Canada, Catalogue No. 12-586-XIE.

Trépanier, J. (2005). "Data Integration Project Plan". Internal document, Statistics Canada.

Yung, W., Cook, K. and Thomas, S. (2004). "Use of GST Data by the Monthly Survey of Manufacturing". *Proceedings of the Section on Survey Research Methods,* American Statistical Association, CD-Rom.

Appendix A : Respondent / Nonrespondent Components at the Data Collection Phase



Appendix B : Respondent / Nonrespondent Components at the Estimation Phase