

## REPORTING RESPONSE RATES IN SURVEYS WITH MIXED COLLECTION MODES

Carlos A. Leon<sup>1</sup>

### ABSTRACT

The increasing use of administrative data in annual and monthly business surveys raises several methodological issues. One of them is how to define response rates when survey and administrative data are combined at the record level. The purpose of this note is to propose a response rate definition for surveys where a mixture of direct collection and administrative data is used.

KEY WORDS: Administrative Data, Business Surveys, Response Rates

### RÉSUMÉ

L'utilisation croissante des données administratives dans les enquêtes-entreprises annuelles et mensuelles soulève plusieurs questions méthodologiques. Une d'entre elles touche la façon de définir les taux de réponse lorsque les données d'enquêtes et administratives sont combinées au niveau de l'enregistrement. L'objectif de cette note est de proposer une définition pour les enquêtes lorsqu'un mélange de collecte directe et de données administratives est utilisé.

MOTS CLÉS : Données Administratives, enquêtes auprès les entreprises, taux de réponse

## 1 INTRODUCTION

Statistic's Canada Standards and Guidelines for Reporting Nonresponse Rates were adopted to standardize response rates computations across surveys. They establish response rates definitions, the framework in which they are to be used and provide detailed guidelines in their application for the purpose of comparing surveys and analysing trends in data collection and respondent behaviour. The existing Guidelines were designed for censuses and sample surveys which are based on direct data collection from respondents and they do not explicitly apply to surveys or portions of a survey that are based on administrative records. Recently, as more and more of our surveys use administrative records, there has been an increasing need for extending the Guidelines to the aforementioned cases.

A good example of this is given by the Unified Enterprise Survey (UES), which integrates several annual business surveys. The survey was created within the Project to Improve Provincial Economic Statistics (PIPES) in order to become the vehicle for producing reliable annual estimates for all industries at the provincial and industrial levels. The UES uses tax replacement (TR) for a certain number of units in the main sample. Starting in 2005 a certain number of units within this TR sub sample were sent a questionnaire covering only the characteristic variables, the remaining revenue variables being retrieved from tax records. Thus, depending on the collection mode used, we may have three possible types of units in our sample.

The current Guidelines apply only to the direct data collection units, but the tax replaced units can also be handled in a similar way (see Trépanier *et al.* 2005). However, a new approach is necessary for the units made up from a mixture of

---

<sup>1</sup>Carlos Leon, Statistics Canada Tunney's Pasture, R. H. Coats Building, Ottawa (Ontario), K1A 0T6, Canada,  
carlos.leon@statcan.ca

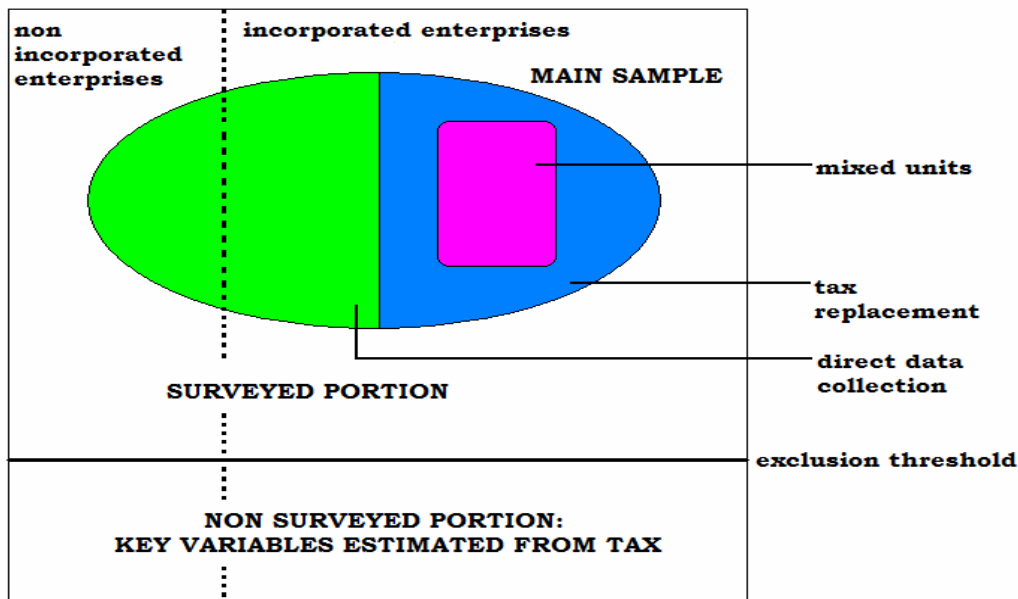
direct collection and administrative data. In what follows, we give a rapid overview of the UES' relevant features and then propose a method for computing response rates for all three types of units; our analysis will be focused on the estimation phase.

### 1.1 Overview of the UES.

The Unified Enterprise Survey integrates several annual business surveys. To reduce response burden on small enterprises, UES uses an exclusion threshold based on revenue (known as the Royce-Maranda threshold) to partition the population into two pieces, i.e. a first portion for which units will be eligible to receive a questionnaire and a second portion for which units will not be eligible to receive a questionnaire. The first portion, known as the take-some portion, is estimated with survey and if applicable, administrative data. The second portion, known as the take-none portion is estimated with administrative data only.

To further reduce response burden, a sub-sample among the take-some, known as the TR portion, is used for tax replacement. No questionnaire is sent to these units and all survey variables are either directly substituted (for revenue variables) or estimated (for characteristic variables) using tax data. Starting in 2005 a certain number of units within the TR sample were sent a questionnaire covering only the characteristic variables, the remaining (revenue) variables being retrieved from tax records. Thus, depending on the collection mode used, we may have three possible types of units in our sample: those obtained from direct data collection, those obtained entirely from tax data and finally those obtained from a mixture of direct data collection and tax data.

The following diagram summarizes the UES sampling strategy for 2005



### 1.2 Extension to characteristic surveys: theory

As mentioned in the introduction, the guidelines allows to compute response rates when the units are either of the direct data collection type or obtained entirely from tax records, but does not apply to the case when a unit is a mixture of direct collection and tax data.

A simple approach to force them into the existing scheme could be to classify such units as direct data collection or extracted from tax data depending on which of the two portions is considered the most important for the survey.

However, doing so would discard valuable information about the portion that is not used. In fact since we are (artificially) creating non-response for one collection mode, it could introduce bias and increase the variance in the estimation of the corresponding probabilities of response

A natural way of handling the mixed unit  $u_i$  is to consider it as a pair  $(u_i^c, u_i^a)$  of two separate sub-units, the first one containing the characteristic variables obtained from direct data collection and the other containing the remaining variables extracted from administrative data. Within the extended classification scheme each of these sub-units would pass through the sieve on its own, one along the direct collection path and the other along the administrative path, eventually contributing to response or non-response in the respective path: the resulting response rates for direct collection and tax data would hence remain basically as they are in the existing scheme. So far this approach seems correct, but we will have to solve a few problems when trying to define the various combined response rates.

Consider a partition  $\{C, A, M\}$  of all the units into three mutually exclusive sets that contain respectively the three types of units: collected directly, from administrative data and mixed type. To avoid unnecessary technical pitfalls we will suppose all units are either in scope or out of scope and that for every mixed unit both components are available.

Let the number of inscope direct collection units be  $N_C = \sum_{i \in C} inscope_i$  where  $inscope_i = resp_i + nresp_i$  are 0-1 variables. Similarly, let  $N_A = \sum_{i \in A} inscope_i$  be the number of (estimated) inscope administrative units. With

$$R_C = \sum_{i \in C} resp_i, \quad R_A = \sum_{i \in A} resp_i$$

the response rates for direct collection and administrative data take the form

$$\frac{R_C}{N_C}, \frac{R_A}{N_A} \quad (1)$$

For the mixed units let's first determine which ones are inscope; recall that to do so it is necessary to consider the pair  $(u_i^c, u_i^a)$ :  $N_M = \sum_{i \in M} inscope_i$

In the UES practice for example, because the direct collection portion is expected to give a more precise idea of the unit's status, we would simply look at the status of the first component  $u_i^c$ . Instead of obtaining the response rate for the mixed units using dichotomic variables and counts, we will derive them from the response status of its components. Let the total response for direct collection and administrative data being respectively  $R_M^C = \sum_{i \in M} resp(u_i^c)$ ,  $R_M^A = \sum_{i \in M} resp(u_i^a)$  where both sums extend over inscope units as defined above. The

corresponding rates for mixed units are

$$\frac{R_M^C}{N_M} = \frac{\sum_{\{i \in M : inscope_i = 1\}} resp(u_i^c)}{\sum_{i \in M} inscope_i} \quad \frac{R_M^A}{N_M} = \frac{\sum_{\{i \in M : inscope_i = 1\}} resp(u_i^a)}{\sum_{i \in M} inscope_i} \quad (2)$$

For defining the response status of the mixed unit we now introduce  $resp_i = f(resp(u_i^c), resp(u_i^a))$  where  $f$  satisfies the following natural conditions:  $0 \leq f \leq 1, f(0,0) = 0, f(1,1) = 1$

Having defined the response, the total response is simply  $R_M = \sum_{i \in M} resp_i$  and the combined response rate for mixed units can be written as

$$\frac{R_M}{N_M} = \frac{\sum_{\{i \in M: inscope_i=1\}} f(\text{resp}(u_i^c), \text{resp}(u_i^a))}{\sum_{i \in M} inscope_i} \quad (3)$$

Now, when the three sources are accounted for direct collection, admin data and global response rates become respectively:

$$\frac{R_C + R_M^C}{N_C + N_M}, \frac{R_A + R_M^A}{N_A + N_M} \quad \text{and} \quad \frac{R_C + R_A + R_M}{N_C + N_A + N_M} \quad (5)$$

So far we have left the function  $f$  that defines the response value for mixed units in general form; hence the rates (3) and (5) are defined at that level of generality. But in practice there are only a few natural candidates one needs to consider; let us study the following two possibilities:

$$f_1(a, b) = \lambda(a \vee b) + (1 - \lambda)(a \wedge b), \quad 0 \leq \lambda \leq 1 \quad f_2(a, b) = \theta a + (1 - \theta)b, \quad 0 \leq \theta \leq 1$$

When comparing the two functions,  $f_1$  can be viewed as  $f_2$  applied on the symmetrized pair  $(a \vee b, a \wedge b)$ . If  $\Delta(a, b)$  denotes their difference, we have:

$$\sum_{\{i \in M: inscope_i=1\}} \Delta(a_i, b_i) = (\lambda - \theta) \sum_{\substack{\{i \in M: inscope_i=1\} \\ a_i \neq b_i}} a_i + (\lambda + \theta - 1) \sum_{\substack{\{i \in M: inscope_i=1\} \\ b_i \neq a_i}} b_i$$

and the equation  $\sum \Delta(a_i, b_i) = 0$  has always a unique solution in  $\lambda$  when  $\theta$  is fixed, but the contrary does not hold. Another reason that makes the symmetrized form  $f_1$  more appealing is that it accounts for the interaction between the two components, whereas  $f_2$  only accounts for the marginals. Considering the above we would advocate the use of  $f_1$  but again there could be cases where  $f_2$  or another function might be preferable.

### 1.3 Extension to characteristic surveys: updating the categorization diagram

From the above it should be now clear that in order to compute response rates for the characteristic units according to our program we need two ingredients: on one hand a categorization scheme and, on the other hand, a function  $f$  that will allow us to account for the two separate components and determine the response value of the inscope units. The existing classification can be used for the first, but some important modifications will be required in order to identify in scope units in a coherent way. The key idea used to perform this modification is that for determining the in scope status the entire unit is needed whereas the response value is determined by the response status of its collection and admin components. We have only included the categories needed for the response rates computations, and further details can be added if necessary.

The main current classification is presented in Appendix A with boxes delimited by a solid line, whereas the extension to administrative units from Trépanier *et al.* 2005 and the proposed additions are delimited with dashed lines. The definitions are given for the main categories only, the corresponding definitions for their dichotomic counterparts being obtained in an obvious way.

Total Units:	Composed of any unit included in the census or sample survey that is meant by design to be observed either by direct data collection or by data extraction from an administrative file or using a mixture of direct collection and administrative data.
Resolved:	Unit or component whose status has been resolved by the end of the data gathering period has either belonging or not belonging to the target universe for the survey.
In-scope:	Resolved unit determined to belong to the target universe for the survey
Extracted:	Unit or component that was meant by design to be obtained from an administrative file and for which the extraction was successful.
Reporting:	Unit or component estimated in-scope for which sufficient reported tax information is obtained
Responding:	In-scope unit or component which is deemed to have responded.
Estimated In-scope:	Unresolved unit estimated to be in-scope for the survey.

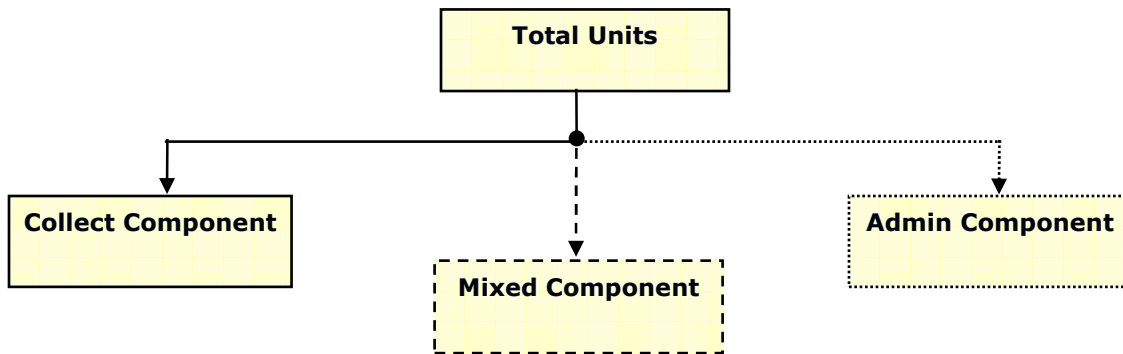
#### 1.4 Conclusion

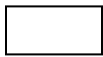
The increasing use of administrative data in our surveys has created the need for new quality indicators, and in particular response and nonresponse rates. We have proposed a method that allows to handle our current business surveys and specially those where collected and admin data are combined at the record level. This method should cover our needs until even more audacious ways of integrating administrative data in our surveys are developed.

#### REFERENCES

- Trépanier, J., Julien, C., Kovar, J. (2005). *“Reporting response Rates when Survey and Administrative Data are Combined”*. Federal Committee on Statistical Methodology Research Conference.
- Statistics Canada (2001). *“Standards and Guidelines for Reporting of Nonresponse Rates: Definitions, Framework and Detailed Guidelines”*. Internal document, Statistics Canada.

## Appendix A: Classification Diagram at the Estimation Phase

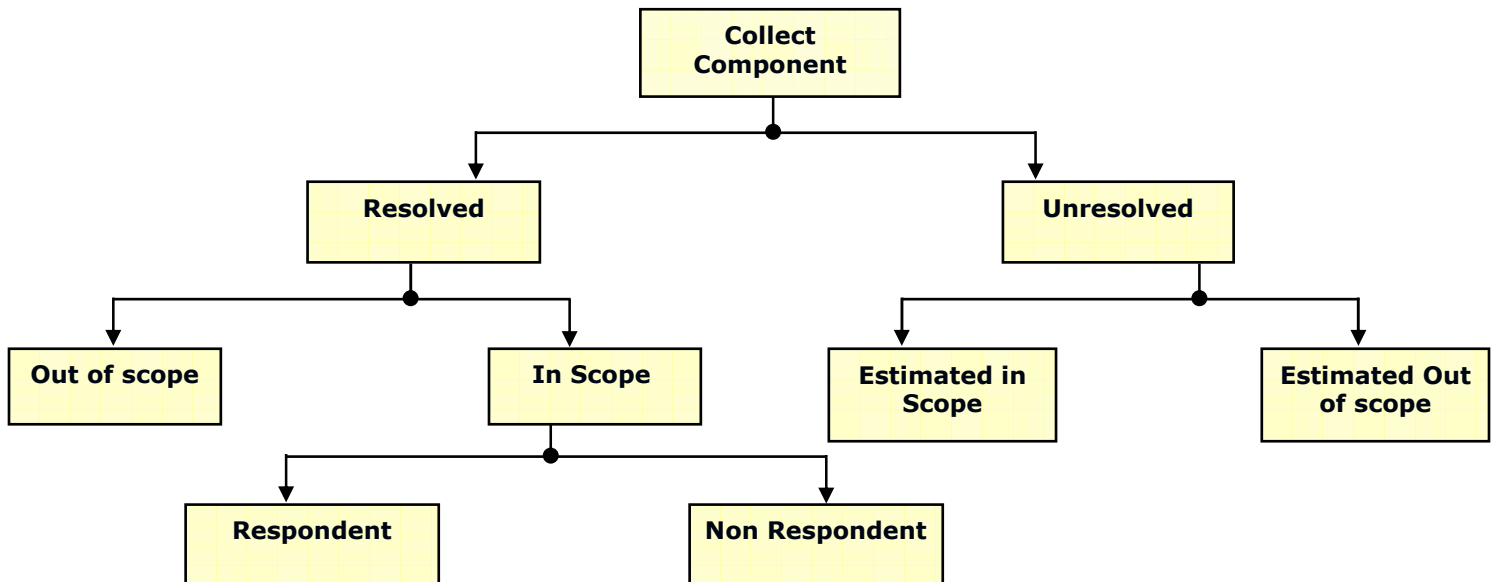


 In the Standards and Guidelines

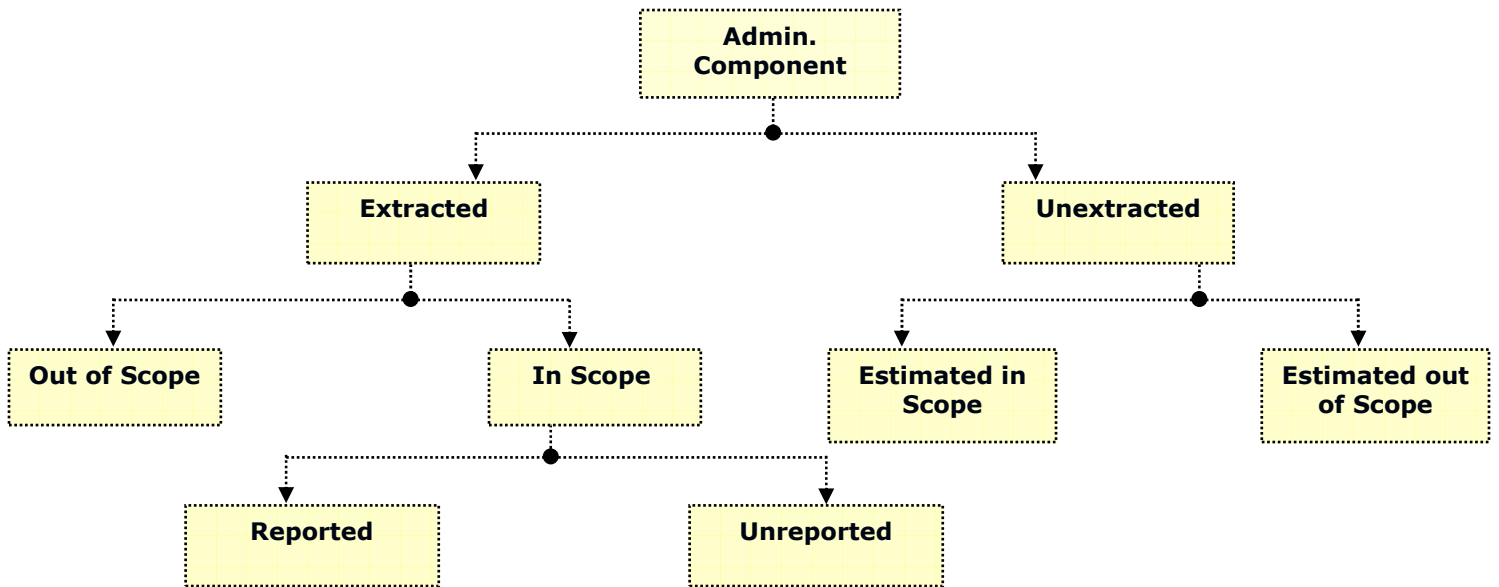
 Extension to Administrative Units

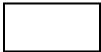


 Proposed Addition

**Appendix A: Classification Diagram for Collected Units at the Estimation Phase**

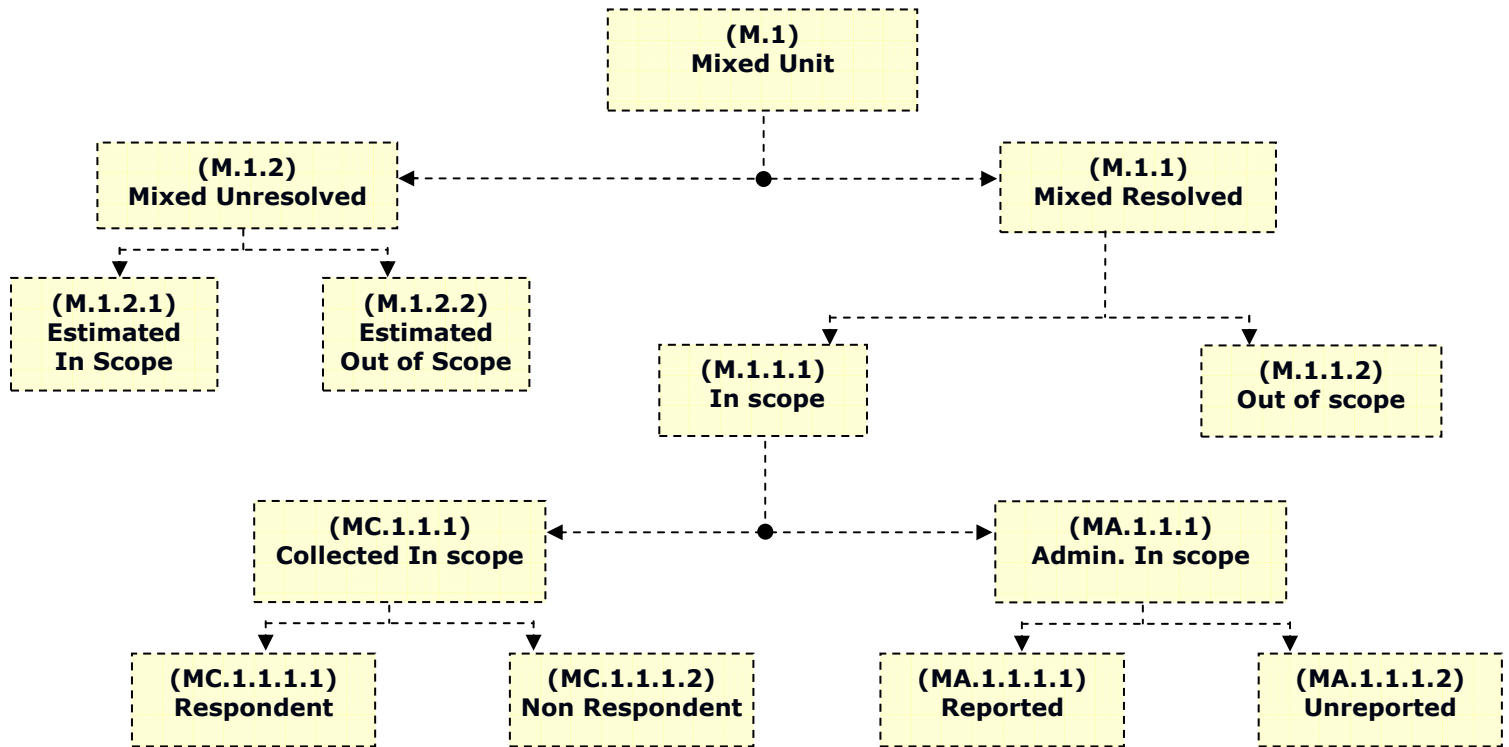


## Classification Diagram for Admin Units at the Estimation Phase



-  In the Standards and Guidelines
-  Extension to Administrative Units
-  Proposed Addition

## Appendix A: Classification Diagram for Mixed Units at the Estimation Phase



- In the Standards and Guidelines
- Extension to Administrative Units
- Proposed Addition