



Catalogue no. 11-522-XIE

**Statistics Canada International Symposium
Series - Proceedings**

**Symposium 2003: Challenges
in Survey Taking for the Next
Decade**

2003



Proceedings of Statistics Canada Symposium 2003
Challenges in Survey Taking for the Next Decade

WEIGHTING METHOD USED IN THE FRENCH “CONSTRUCTION DES IDENTITÉS” SURVEY

Pascal Ardilly¹

ABSTRACT

The “Construction des identités” survey involved 13,500 individuals from multiple sources and over-represented the populations of disabled people, immigrants and children of immigrants on the date of the census. When the selected individual was not there, another person was randomly sampled from the household that the enumerator visited on the collection date. The weighting system presented serious problems due to overlapping survey frames and to the process of randomly selecting some individuals. In particular, there were difficulties relating to the unavailability of certain pieces of information used in determining an “unbiased” weight.

KEYWORDS: Missing Information; Multiple Frames, Over-Representation; Two-Phase Survey; Weighting.

1. OBJECTIVE OF THE SURVEY

The objective of the “Construction des identités” survey, conducted in France in March/April 2003, was to describe and analyze the types of social bonds that enable individuals to integrate into French society at the beginning of the 21st century. It involved 13,500 individuals who were at least 18 years of age from multiple sources, and over-represented three populations that are more likely to have difficulties integrating: immigrants, people born to immigrant parents and persons under the age of 60 with disabilities. However, the issue of identity, understood in the broad sense as the way individuals build a place for themselves in society that allows them to both integrate and affirm their own individuality, arises also for people who do not originate from outside France and who have no physical or mental disabilities; hence the reason why over half the sample involved people who *a priori* do not have any of the preceding three characteristics.

2. BUILDING THE SAMPLE

2.1 Initial selections in the survey frames

The first part of the sample comes from a very large survey (about 400,000 individuals) called “Vie quotidienne et santé” (VQS, daily life and health), an addition to the 1999 census, that targeted people under 60 with disabilities. It was a stratified, areal survey that involved whole enumerator sectors previously sampled by simple random sampling within delegate zones,² themselves resulting from a survey. These delegate zones were selected proportionate to their size in number of individuals.

In “Identités”, the disabled persons sample was a second selection phase within the VQS sample: using the VQS delegate zones that were previously subsampled³, a systematic selection was done of 2040 disabled individuals from all zones. Then, the size of the sample cutting across each delegate zone was reviewed. Where there were 6 or fewer people in a given zone, they were all surveyed; otherwise another random sampling was done so as not to exceed the limit of 6 surveys per zone. In the end, 1215 individuals made up sample S1. At first glance, the final selection probabilities of the individuals should be close, since, on one hand, the delegate zones were initially drawn

¹ Pascal Ardilly, INSEE, 18 Bd Adolphe Pinard 75675 Paris Cedex 14, France (pascal.ardilly@insee.fr).

² A census delegate has a jurisdiction that includes approximately 40 enumerators.

³ For 2 reasons: first, we wanted to limit the zone sample size in “Identités” and second, VQS over-represented some regions, and we wanted to restore the balance of the sampling rates for “Identités”.

proportionate to their size, and on the other, the sampling in the second stage was designed to ensure constant allocation by zone, or less than the limit of six in case of insufficient reserve.

A second part of the sample involved two populations identified during the “Etude de l’histoire familiale” (EHF, family history study) survey, also very large and an addition to the 1999 census, namely immigrants and children of immigrants. Like the VQS, the EHF survey was stratified, areal and involved whole enumerator sectors in the same delegate zones as the VQS, but nevertheless different than the enumerator sectors used by the VQS. The EHF sampling design was also developed to over-represent women.

The “Identité” sampling of immigrants and children of immigrants was considered a second selection phase in the EHF and was done in the delegate zones selected for sample S1 (to limit the number of delegate zones called on and, as such, the number of enumerators). We therefore identified geographically broad delegate zones and small zones⁴. The final processing was organized into two batches.

First, we considered the set consisting of small zones and large zones with 3 communes or less, and were interested in the immigrant population residing there. We then performed a systematic selection within this population. We then used this principle for sampling disabled people: we looked at the number sampled per delegate zone, and if it was below a certain limit, we kept everyone; otherwise we performed another systematic, conveniently sized sampling (equalling 10 individuals).

Second, we considered the large zones consisting of more than 3 communes. We started by selecting 3 communes in each zone, proportionate to their size, and then commune by commune sampled a given number of individuals (5 people) – unless the available number was insufficient (which was relatively common with the “rare” populations we were concerned with), in which case we surveyed all available people.

Finally, we did a simple random sampling of women to “erase” the over-representation introduced by the first EHF phase.

We ended up with a sample of 2025 immigrants, designated S2. The same entire procedure was used with the children of immigrants, resulting in sample S3 of 2025 people as well. For both S2 and S3, with the same argument as for S1, we should have had a relatively limited weight spread, because, if we disregard the cases in which the selectable reserve was below the limit, the selections involved were multi-stage and proportionate to the size at each stage, except in the last one where the number selected was constant.

A fourth sample, designated S4, comprising 7830 individuals, was drawn directly from the 1999 census files, independently of the previous samplings except if it was located entirely in the delegate zones previously sampled for S1, S2 and S3 but avoided the enumerator sectors that the VQS and EHF surveys focused on (these sectors were determined using an equal probabilities selection in the delegate zone). We identified new small and large zones, sampled communes in the large zones if they had 4 or more communes, and selected individuals by simple random sampling in the selected zones or communes, depending on the case. The allocations were agreed to in advance, specifically 21 individuals in the small zones and 8 individuals in each of the 3 selected communes for the large zones. If there were fewer than 3 communes in a large zone, we tried to arrive at a size of 21 people selected in the zone, to limit the weight spread.

Finally, in a specific frame completely unconnected with the previous ones, we sampled 405 new dwellings in the chosen delegate zones, again according to a complex design, to take into account any construction completed after 1999. As such, we obtained sample S5.

The overall sample was therefore the combination of five samples (S1 to S5) from four different sources and comprising 13,500 individuals. It reflected over-representation of disabled persons under 60 years of age, immigrants and children of immigrants. Each sampling was stratified, multi-stage, and involved a complex sampling design where we tried to limit the weight spread of individuals. At the end of the process, we identified and eliminated duplications; in other words, we kept only one individual in households that were called on at least twice

⁴ In the zone survey frames, there is a total area indicator.

during the S1 to S4 selections. These duplications were not anecdotal since the populations involved in S1, S2 and S3 exhibited relatively natural clustering.

2.2 Selection of the individuals questioned

Since the selected individuals were identified with their characteristics as of the date of the census, particularly their residence address where they were living when the census was conducted, there were of course many cases where the selected individuals were not there, most often because they had moved away after March 1999, the census date. Since there was no follow-up on these individuals at their new addresses, the procedure in such cases was to randomly sample a person from those in the household visited by the enumerator on the collection date.

3 . WEIGHTING DIFFICULTES ENCOUNTERED

The sampling design used created a number of obstacles to a simple weighting calculation that can potentially cause a certain loss of effectiveness:

- regarding the selection process itself:
 - 1) As a matter of principle, we should emphasize the complexity of the sampling, even though we limited ourselves to a specific population. For S1, S2 and S3, the process was a two-phase survey in which each phase arose from an unequal probabilities and multi-stage design. The allocations eventually selected in the various geographic units – delegate zones and communes – were not only tied to the unit in question but were also contingent upon the entire sampling process: for example, using systematic selection of a given size out of a set of delegate zones (of varying sizes) made the allocation in each one totally dependent on the entire list of zones sampled. This element of complexity appeared indispensable from the moment when the numbers per sample were set and we necessarily had to proceed with a two-phase selection process in which the first-phase sample varied in size.
 - 2) In practice, sample determination sometimes strayed from theory to take into account field requirements that did not always leave traces or due to computer problems probably pernicious in origin, in conjunction with insufficient documentation on building survey frames. In particular, reviewing the final samples of the VQS and EHF surveys revealed several “curiosities” that made the weight calculations fragile, since, for example, the reported population in some delegate zones seemed to deviate more from the standard size (15,000 inhabitants) than we would at first have thought, or because the very number of delegate zones cut across in an area was not known for certain. Another example: we had to deal with several delegate zones that gave rise, after some routine eliminations (previous to each selection, we removed the samples of dwellings that were called on in the previous surveys) to a selectable number that was lower than the expected number.
 - 3) The gross weight spread, sample by sample, eventually appeared more significant than expected, since it varied in a 1-to-10 ratio if we excluded several extreme weights. This is in part related to the problems indicated in the previous point, particularly to the fact that the population sizes of the delegate zones used to calculate the weights were noticeably more spread out than anticipated: the initial zone subsampling at the start of the second phase was based on the idea that the delegate zones should be calibrated in order to be close to an ideal average size, but this proved to be counter to current practice, and a weighting inequality resulted from this misinterpretation. Another explanation involves the allocation spread by zone or by commune: in fact, consistent weighting was not achieved unless the allocation by zone or by commune was constant, which was not at all guaranteed, being relatively small populations. The threshold effect had therefore not been fully assessed either. Finally, it must be remembered that delegate zones were selected in proportion to the total size in number of inhabitants but that the selections in the final stages used the sizes of the interest populations (disabled / immigrants / children of immigrants): however, the relative weight of these populations varied substantially from one commune to the next.

The table below provides statistics on the gross weights obtained from the initial selections:

| | Minimum | Decile 1 | Average | Decile 9 | Maximum |
|-----------------------------|---------|----------|---------|----------|---------|
| Disabled (S1) | 253 | 560 | 1,403 | 2,523 | 6,886 |
| Immigrants (S2) | 256 | 975 | 2,640 | 5,279 | 11,976 |
| Children of Immigrants (S3) | 252 | 952 | 2,332 | 3,933 | 19,270 |
| Other (S4) | 927 | 3,371 | 5,628 | 7,993 | 21,451 |

- regarding the operations following the selection:
 - 1) A disabled individual, immigrant or child of immigrants can be selected through several survey frames: in fact, he/she may come from one of samples S1, S2 or S3, but also from S4. If there are several combined characteristics, for example, a person who is both an immigrant and a child of immigrants, he/she can be selected through 3 frames, even 4 (since a person may be an immigrant, a child of an immigrant, disabled and under 60 all at the same time).
 - 2) Let us consider an individual who is disabled or an immigrant or the child of immigrants. If he/she is selected through either sample S1, S2 or S3, it is possible to calculate a conditional probability of selection, as we did in the multi-phase selections because we mastered the series of steps that enabled us to select it. This selection probability is not “the” selection probability used in the Horvitz-Thompson estimator, but it nevertheless provided an unbiased estimator. His/her classic probability of selection can also be calculated *a priori* in S4, because this time it is a relatively simple multi-stage design in a single phase. On the other hand, if the individual under consideration comes from S4, it is very much wishful thinking to hope to obtain “something” nearing a selection probability relating to designs leading to S1, S2 or S3, given the complexity of these designs and particularly, as it seems to us, the need to use a conditional approach. Therefore, there was great difficulty in calculating a gross weighting for people who were disabled, immigrants or children of immigrants if they came from an S4 sampling.
 - 3) The individual’s status – which can take 4 forms: disabled, immigrant, child of immigrants or “other” – was unknown at selection time and was known only after the collection phase. We therefore had to reconstruct the March 1999 situation by adequately questioning individuals (questions were prepared for this purpose), which necessarily resulted in observation errors. In fact, it is easy to imagine that there is a certain vagueness around what constitutes a disability (this is less the case for the immigrant and children-of-immigrants statuses, although inconsistencies in the answers given in a four-year interval may also exist in relatively factual items). The weight calculation, which essentially depends on status, can therefore only be done late.
 - 4) The process of randomly selecting an individual in a household when the originally designated individual had left the residence confused the issue: in fact, the weighting of the individual ultimately questioned – whether enumerated or not – had to depend to a certain extent on the selection probabilities of the people who left the household after the census. Unfortunately, it was virtually impossible to obtain any information about the 1999 status of these individuals. Moreover, if the individual ultimately questioned was there at the time of the census but designated as resulting from a random selection within the household, he/she also had an initial selection probability that was not reasonably calculable (in fact, this individual could have been selected at least through S4, perhaps even through S1, S2 or S3). In the case of scattered households (there were about 40), the individual was virtually inextricable.

This type of difficulty alone was one of the most detrimental to the process; in a way, it “destroyed” all efforts in the previous steps to assign an initial weight that was relatively consistent with the sampling adopted.

- 5) The selection of new dwellings was based on a survey frame that placed the dwellings at the communal level only. However, the selections had to be concentrated in certain delegate zones, as explained in 1), but numerous delegate zones were infra-communal. This information mismatch could only lead to very simplifying calculations that did not reflect the reality of the sampling design.

4. PROCESS USED FOR OBTAINING A WEIGHTING

4.1 Selection of individuals in their respective frames

We therefore did not attempt to calculate a selection probability that explicitly took into account the various sampling designs connected with the different frames through which an individual could be selected. However, we did calculate a gross weight associated with each individual selected from each of samples S1 to S4, a preliminary step that seemed indispensable! We then used the following simple principle: if S_a and S_b are 2 independent samples drawn from the same population, with the respective selection probability sets $\Pi_k^{(a)}$ and $\Pi_k^{(b)}$ for each individual k , a natural unbiased estimator of the total is:

$$\theta \cdot \sum_{k \in S_a} \frac{Y_k}{\Pi_k^{(a)}} + (1 - \theta) \sum_{k \in S_b} \frac{Y_k}{\Pi_k^{(b)}} \quad (1)$$

where θ is a real number between 0 and 1. The optimal value of θ is connected with the variance of each of the respective estimators obtained on S_a and S_b . The ratio of the size of sample S_a to the total size of sample $S_a \cup S_b$ is a reasonable choice. This approach easily extends to the case of 3 or more samples.

Thus, for each of the three statuses, “exclusive” disabled / “exclusive” immigrant / “exclusive” child of immigrants, and for each possible overlap between statuses (making a total of 7 configurations), we obtained weights by concatenating the following steps:

- identify samples likely to contain individuals with the given status;
- calculate coefficient θ (or several coefficients) proportionate to the size of the sample associated with θ . Where S4 is involved, in other words always, the size considered is that fraction of S4 that has the status in question;
- determine the selection probabilities using the probabilities calculated from the initial sampling design and dividing them by the coefficient θ , for the fraction of the sample concerned by the status

We did not really obtain selection probabilities, but rather pseudo-probabilities, the inverse of which represented an acceptable weight. This operation assumed that we could always obtain 1999 status information, because we had to be able to assign a status to every individual selected. As previously indicated, observation errors were a not insignificant source of bias.

Let us look at an example: we identify people who were “exclusively” disabled in 1999 (according to their statements): some are from S1, others from S4, but they cannot come from S2 or S3 (otherwise they would not be exclusively disabled). Those from S1 have a weight w_k calculated from the S1 sampling design, which involves, say, 2,000 individuals. Those from S4 have a weight w_k calculated from the S4 sampling design and total, say, 1,000 individuals. Therefore, we would simply give each S1 individual the weight $w_k * 2/3$ and to each of the 1,000 S4 individuals, the weight $w_k / 3$.

4.2 Selection of the individual surveyed in the household

We then had to go from these pseudo-selection probabilities of the initially selected individuals to the weight associated with each individual ultimately questioned. To do so, we considered the household. The set of individuals

present both at the time of the 1999 census and on the date of the survey, we labelled A. The set of individuals present at the time of the 1999 census but who had permanently left the household by the date of the survey, we labelled B. Finally, the set of individuals in the household who were not enumerated in 1999 in the designated dwelling, we labelled C. Within household \mathbf{m} , we labelled Π_k^I the pseudo-selection probability of individual \mathbf{k} (obtained according to 4.1) and Π_k^F the final selection probability of individual \mathbf{k} , as it resulted from the field process.

We disregarded the possibility of two separate individuals being selected in the same household⁵. In this case, we had:

- For every \mathbf{k} in A:
$$\Pi_k^F = \Pi_k^I + \frac{1}{N_m} \cdot \sum_{j \in B} \Pi_j^I$$
- For every \mathbf{k} in C:
$$\Pi_k^F = \frac{1}{N_m} \cdot \sum_{j \in B} \Pi_j^I$$

where N_m is the size of household \mathbf{m} on the date of the survey (number of individuals in the field). We particularly observed that if B was empty and C was not, there was a bias because the new individuals were not covered by the methodology used.

As previously regretted, we knew absolutely nothing about the Π_k^I of the B individuals, and elected to impute a value equal to the average selection probability $\bar{\Pi}$. This average value – numerically close to 4500 – was obtained using

$$\sum_c \frac{\hat{N}_c}{\hat{N}} \bar{\Pi}_c = \bar{\Pi} \tag{2}$$

where \mathbf{c} represents the four possible statuses, $\bar{\Pi}_c$ is the arithmetic mean of the Π_k^I of all individuals selected whose status \mathbf{c} we knew (necessarily individuals from group A initially designated randomly), and \hat{N}_c estimates the size of the population of status \mathbf{c} (obtained using the probabilities Π_k^I). The calculations were done out of all of France. Only those individuals selected in households that had not lost an individual since the census were excluded from this imputation. This approach required knowing the cardinality of B, which we found out through a question specifically introduced to determine whether each individual making up the household on the date of the survey was or was not present in that same dwelling at the time of the census (since we knew the household size at the time of the census, we deduced the size of B by subtraction – however, it must be recognized that we encountered inconsistencies here, such as sizes of households enumerated in March 1999 lower than the cardinality of A).

The problem of not knowing the Π_j^I of individuals \mathbf{j} of A selected randomly in the household was handled in the following manner – on the understanding that one believes in the relevance of these imputations, which are subject to discussion and claim to give only rational orders of magnitude:

We labelled \mathbf{k} the initially designated individual (necessarily belonging to B):

- \mathbf{j} is neither disabled, an immigrant, nor child of immigrants: he/she can come only from S4 and we recalculated a selection probability using the S4 sampling design that we designated as the value of Π_j^I .

⁵ It was low but not nil. Technically, we were not able to calculate it precisely, and the approximation was very complex.

- **j** is either disabled, an immigrant, or child of immigrants, and **k** happened to have the same 1999 status as **j**, and that **k** came from S1 or S2 or S3. We took Π_k^I as the value of Π_j^I (this did not work well for cases of overlapping status – for example disabled immigrants – because we had only incomplete information on **k**, but these cases must be rare, and to avoid too many complications, it was not taken into account).
- **j** is either disabled, an immigrant, a child of an immigrant, or a combination of two or three of these statuses, and **k** happened to come from S4. We took $\bar{\Pi}_c$ as the value of Π_j^I where c represented the status of **j**.
- other cases: we imputed $\bar{\Pi}$.

4.3 New dwellings

A word on new dwellings, to point out that the problem raised in 3 was bluntly resolved: an estimate of new construction was done on a France-wide basis, and we assigned a weight equal to the inverse of the sampling rate, in other words consistently equal to 3162.

4.4 Adjustments and correction of nonresponse

Approximately 8,500 individuals responded to the survey. We assumed that the random selection of an individual in a household was never triggered by nonresponse on the part of the initially designated individual. Using the gross weights obtained using the previous process, we used the CALMAR software to perform a calibration of the external data that could both correct the total nonresponse bias and reduce the sampling variance.

Table 1: Comparison between the margins drawn from the sample (with initial weighting) and the margins in the population (calibration margins)

| | Sample Margin | Population Margin | % Sample | % Population |
|---|---------------|-------------------|----------|--------------|
| Born abroad | | | | |
| Yes | 5251.84 | 5629.50 | 11.66 | 12.50 |
| No | 39784.16 | 39406.50 | 88.34 | 87.50 |
| Number of people in the household | | | | |
| 1 | 8039.00 | 7701.16 | 17.85 | 17.10 |
| 2 | 16149.24 | 15447.35 | 35.86 | 34.30 |
| 3 | 7826.20 | 8827.06 | 17.38 | 19.60 |
| 4 | 7933.87 | 8106.48 | 17.62 | 18.00 |
| 5 and + | 5087.69 | 4953.96 | 11.30 | 11.00 |
| Occupation of the person questioned | | | | |
| Salaried employee | 21575.33 | 21076.85 | 47.91 | 46.80 |
| Self-employed | 2328.94 | 2522.02 | 5.17 | 5.60 |
| Unemployed | 2768.50 | 2296.84 | 6.15 | 5.10 |
| Student | 2550.87 | 3107.48 | 5.66 | 6.90 |
| Retired | 9378.48 | 8151.52 | 20.82 | 18.10 |
| Out of business | 1682.44 | 2386.91 | 3.74 | 5.30 |
| Not in the labour force | 4751.44 | 5494.39 | 10.55 | 12.20 |
| Age group by gender | | | | |
| Male under 29 | 3329.21 | 4503.60 | 7.39 | 10.00 |
| Male 29 to 38 | 3830.01 | 4233.38 | 8.50 | 9.40 |
| Male 39 to 48 | 4160.26 | 4098.28 | 9.24 | 9.10 |
| Male 49 to 58 | 3671.00 | 3692.95 | 8.15 | 8.20 |
| Male 59 to 68 | 2484.00 | 2431.94 | 5.52 | 5.40 |
| Male 69 and older | 2993.59 | 2657.12 | 6.65 | 5.90 |
| Female under 29 | 4085.82 | 4458.56 | 9.07 | 9.90 |
| Female 29 to 38 | 4427.67 | 4278.42 | 9.83 | 9.50 |
| Female 39 to 48 | 4811.55 | 4233.38 | 10.68 | 9.40 |
| Female 49 to 58 | 4197.02 | 3783.02 | 9.32 | 8.40 |
| Female 59 to 68 | 3061.22 | 2702.16 | 6.80 | 6.00 |
| Female 69 and older | 3984.64 | 3963.17 | 8.85 | 8.80 |
| Type of settlement and commune | | | | |
| Rural commune | 11724.21 | 12114.68 | 26.03 | 26.90 |
| Individual, com. with under 50,000 people | 8465.82 | 6980.58 | 18.80 | 15.50 |
| Collective, com. with under 50,000 people | 2604.18 | 3332.66 | 5.78 | 7.40 |
| Individual, com. with 50, to 200,000 people | 3796.75 | 2882.30 | 8.43 | 6.40 |
| Collective, com. with 50, to 200,000 people | 2741.47 | 2792.23 | 6.09 | 6.20 |
| Individual, com. with 200,000 people and + | 5565.80 | 3873.10 | 12.36 | 8.60 |
| Collective, com. with 200,000 people and + | 5441.11 | 5629.50 | 12.08 | 12.50 |
| Individual, in the Paris conurbation | 1590.92 | 1891.51 | 3.53 | 4.20 |
| Collective, in the Paris conurbation | 1808.41 | 3873.10 | 4.02 | 8.60 |
| "Intracity" Paris | 1297.31 | 1666.33 | 2.88 | 3.70 |

5. CONCLUSION

What was interesting – and useful – in this experience was as much the type of difficulties encountered as the technical solutions applied.

- 1) There is a general trend, which is not new, to consider that the technical issues around surveys are not fundamental to the information process. At any rate they are not of the sort that affects major decisions establishing the main features of the survey. Naturally, it would be difficult to admit to civil society that the technique is an obstacle that can deprive it of knowledge of some aspects of its own functioning, but in any decision, the risk of quality degradation must also be weighed. Similarly, the argument for simplicity seems to us to be a strong argument: this appears obvious, but in the end all or substantially all the difficulties encountered in assigning an acceptable weighting were only the result of a gap between the constraints of the operation and the capability, in the broad sense, of the team responsible for conducting it.
- 2) The weighting was designed to favour simplicity to the extent possible in the sampling and computer procedures that were models of complexity. Perhaps there was also trauma connected with the manifold frames and the difficulties in the computer processing of the data, but it now seems to us that the weight-share method would probably have been a more thorough route for justifying that we did not calculate the selection probabilities of individuals using the frame(s) in which they appeared but without being sampled there. However, it would not have resolved the other issues.
- 3) In this particular case, the procedures implemented eventually led to an entirely satisfactory weighting system before adjustments, properly respecting the main socio-demographic structures (see table 1). Should we give ourselves full marks? Don't the "acrobatic" treatments have only a low numeric impact in this instance? Is there any chance ... ? At any rate, in the end, this proves all those people right, this time, who believe that all technical problems have a satisfactory solution.

REFERENCES

- Lavallée, P. (2003), *Le sondage indirect ou la méthode généralisée de partage des poids*, Ellipses.
- Särndal, C.-E., Swensson, B., Wretman, J. (1992), *Model assisted survey sampling*, New York : Springer-Verlag.