Component of Statistics Canada Catalogue no. 11-522-X
Statistics Canada's International Symposium Series: Proceedings

Article

Symposium 2008:
Data Collection: Challenges, Achievements and New Directions

Non-response in a Random Digit Dialling Survey: The Experience of the General Social Survey's Cycle 21 (2007)

by Isabelle Marchand, Ryan Chepita, Patrick St-Cyr and Kuawa Williams

2009

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2008
Data Collection: Challenges, Achievements and New Directions

Non-response in a Random Digit Dialling Survey: The Experience of the General Social Survey's Cycle 21 (2007)

Isabelle Marchand, Ryan Chepita, Patrick St-Cyr and Kuawa Williams[1]

Abstract

The growing difficulty of reaching respondents has a general impact on non-response in telephone surveys, especially those that use random digit dialling (RDD), such as the General Social Survey (GSS). The GSS is an annual multipurpose survey with 25,000 respondents. Its aim is to monitor the characteristics of and major changes in Canada's social structure. GSS Cycle 21 (2007) was about the family, social support and retirement. Its target population consisted of persons aged 45 and over living in the 10 Canadian provinces. For more effective coverage, part of the sample was taken from a follow-up with the respondents of GSS Cycle 20 (2006), which was on family transitions. The remainder was a new RDD sample. In this paper, we describe the survey's sampling plan and the random digit dialling method used. Then we discuss the challenges of calculating the non-response rate in an RDD survey that targets a subset of a population, for which the in-scope population must be estimated or modelled. This is done primarily through the use of paradata. The methodology used in GSS Cycle 21 is presented in detail.

Key Words: Random digit dialling survey, Response rate, Unresolved cases, Subpopulation.

1. Introduction

The General Social Survey (GSS) is an annual cross-sectional survey of the population. It was established by Statistics Canada in 1985 to fill certain gaps in Canadian social statistics. In order to respond to a wide variety of social concerns and federal government policies, the survey has two main objectives: (1) collect data to track changes in the social structure and lifestyle of Canadians, and (2) provide immediate information about specific social statistics. To that end, the GSS program involves administering a survey on five major topics. Each topic is covered every five years in an annual survey known as a cycle. This amount of time is considered reasonable for measuring development, since social change occurs over a long period of time. The most recent topics have been family, social networks, time use, victimization and retirement. Each questionnaire in a GSS cycle has three modules: a common set of items that are repeated every five years to capture trend data (core topic), specific content (particular topic) that varies from cycle to cycle to provide snapshot data, and a classification module that is present in each cycle and consists of a set of basic demographic data variables and socio-economic variables. In its initial cycles, the survey had 10,000 respondents. Since Cycle 13 (1999), the sample has consisted of 25,000 people. The survey uses a computer-assisted telephone interviewing (CATI) system with random digit dialling.

GSS Cycle 21 (GSS-21) was about the family, social support and retirement. Its target population consisted of people aged 45 and over living in the 10 Canadian provinces. For more effective coverage, the sampling plan had two components. The first part of the sample was taken from a follow-up of GSS-20 respondents who were potentially eligible for GSS-21; the second part was a new sample selected using RDD techniques.

This paper is structured as follows. The sampling design of GSS-21 and the RDD method used in the survey are described in the next two sections. Section 4 discusses the challenge of calculating the response rate in an RDD survey that targets a subset of the population. Section 5 provides a detailed explanation of the methodology used in GSS-21 to compute the response rate for the RDD component. Section 6 covers the method used to calculate the response rate for the component from GSS-20 and the overall response rate for GSS-21.

[1] Isabelle Marchand, Ryan Chepita, Patrick St-Cyr and Kuawa Williams, Statistics Canada, R. H. Coats Building, 16th Floor, 100 Tunney's Pasture Driveway, Ottawa Ontario, Canada, K1A 0T6

2. Sampling design of GSS-21

The target population of GSS-21 included all persons aged 45 and over living in the 10 Canadian provinces, excluding full-time institutional residents. Two of the sampling plan's requirements were to provide detailed estimates at the national level and urban-rural comparability for each province. To achieve the survey's objectives, the Canadian population was stratified by provinces, which in turn were divided up on the basis of census metropolitan areas (CMAs), forming a CMA group and a non-CMA group. Canada has a total of 27 strata. Within each stratum, the sample was allocated in such a way as to balance the need for provincial and national data. The allocation was performed using Kish's method (Kish, 1976). Each initial size was then inflated to allow for non-response and household eligibility, since only households with at least one member aged 45 or over were eligible for GSS-21. The size was also adjusted for the fact that the sampling unit was a telephone number and not the household.

As mentioned in section 1, part of the sample was taken from a GSS-20 follow-up. Cycle 20 was about family transitions. Its target population consisted of persons aged 15 and over living in the 10 Canadian provinces. That sample was also selected using RDD techniques. The data were collected during 2006, and the response rate was 67.4%. A total of 13,230 GSS-20 respondents were identified as potentially eligible for GSS-21 (i.e. age 43 or over) and were included in the Cycle 21 sample. It was expected that this would yield 10,500 respondents. The rest of the sample, 14,500 respondents, was from a new RDD sample. GSS-21 collection ran from March to December 2007. The annual sample was split into nine monthly samples, each of which had a collection period that extended over a two-month period known as a wave. The sample was distributed in this way to represent seasonal variations in the information.

3. Random digit dialling

3.1 Frame and sample selection

The GSS uses the RDD method, in particular the elimination of non-working banks technique. In Canada, a telephone number consists of 10 digits, containing a three-digit area code, a three-digit prefix and a two-digit bank indicator. Thus, a bank is formed by the first eight digits and contains 100 possible telephone numbers. For example, (613) 951-4703 is part of the 61395147 bank. The sample frame is made up of a list of working banks. By definition, a working bank contains at least one residential telephone number, according to our administrative data. At Statistics Canada (StatCan), InfoDirect data and billing files from various companies are used to identify residential telephone numbers and, consequently, working banks. For sampling purposes, each bank is assigned to a geographic stratum.

The sample is selected by simple random sampling with replacement of banks within each stratum. The last two digits are then randomly generated to produce a 10-digit telephone number. Each telephone number selected is assigned "listed" status if it appears in our telephone listings and "unknown" if it does not.

3.2 Characteristics

With regard to coverage, the sample frame excludes households that have no telephone service. According to the 2007 Residential Telephone Service Survey (RTSS), this proportion of the population is estimated at 0.9%. Households that use cell phones exclusively are also excluded. This proportion of the population is estimated at 6.4%, again according to the 2007 RTSS. Since we are using the non-working banks elimination method, residential numbers that belong to banks falsely identified as invalid are also excluded. That portion is considered negligible.

In recent years, there has been a decline in response rates in telephone interview surveys, especially those using the RDD approach. It is indeed increasingly difficult to reach people and persuade them to take part in a telephone survey. With the RDD approach, there are few auxiliary variables available to optimize the sampling plan and treat non-response. The portability of telephone numbers is eroding the accuracy of geographic information, and in the long term, it may reduce the sampling plan's effectiveness. Nevertheless, the RDD approach has some advantages. It

takes relatively little time to develop and implement an RDD survey. It is also less costly than personal interviewing. From a statistical standpoint, the method has few design effects.

3.3 Collection

Among the telephone numbers selected for the sample, those with "unknown" status undergo an initial sort, referred to as a pre-collection process, by a company outside StatCan. Out-of-service telephone numbers are identified by an automated pre-dialling system. This step cleans up the sample, the process removing about 24% of the initial sample. The telephone numbers are then sent to regular collection. To manage the response burden, a call limit has been in place since 2006. For RDD surveys, a limit of 20 calls is set for numbers whose status is "listed", and a limit of five calls for numbers whose status is "unknown". It should be noted that if a contact is detected within the five-call limit, the limit is reset to 20.

In the random digit dialling approach, a large part of the sample consists of telephone numbers that do not belong to a household. It is estimated that 54% of the telephone numbers sent to the field will reach a household. For Cycle 21, a household was eligible if at least one of its members was 45 years old or above. That proportion was estimated at 64%. Consequently, the expected overall success rate for GSS-21 was 35% (54% x 64%).

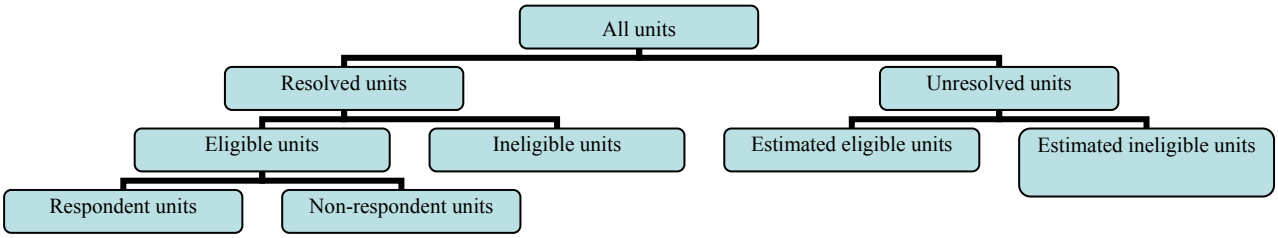
4. Response rate

4.1 Calculation of the response rate

One of the main indicators in assessing the quality of a survey is the response rate. It shows what proportion of the survey is affected by non-response. Non-response can be a major source of non-sampling error if it produces a bias in the estimates. If respondents do not have the same characteristics as non-respondents relative to a variable of interest, and if imputation or reweighting by means of a non-response model or calibration fails to mitigate the difference between them, the estimates or the analytic conclusions may be erroneous. Although there is no direct relationship between bias and a low response rate, the fact remains that the response rate is a key indicator of data quality for users and survey managers.

The aim of the Policy on Informing Users of Data Quality and Methodology is to ensure that users have the information they need to determine how well suited the survey data are to their purpose (Statistics Canada, 2000). To that end, according to Statistics Canada's Standards and Guidelines for Reporting of Non-response Rates (Statistics Canada, 2001), users must be informed of the response rate in the estimation phase. Consequently, units that are unresolved following the collection process must be identified as eligible or ineligible for the survey. Figure 4-1 shows the breakdown of cases at the data estimation stage.

Figure 4-1
Respondent/non-respondent components at the data estimation stage



The response rate provided to users is calculated as follows:

$$\text{Response rate} = \frac{\text{Respondent units}}{\text{Eligible units} + \text{Estimated eligible units}}$$

4.2 Unresolved units in the RDD component of GSS-21

The challenge in calculating the response rate in an RDD survey is the difficulty of determining units' eligibility. As mentioned in section 3.3, only some of the telephone numbers sent to collection are expected to belong to a private residence. The remaining numbers will connect to ineligible units or will be classified as unresolved units. Ineligible units are telephone numbers that are out of service or assigned to businesses, cell phones and so on. Unresolved units are telephone numbers that are busy, are not answered, reach fax machines (that might be in a private residence) or answering machines that provide no clear indication of being in a private residence, or have some technical problem.

Various factors contribute to an increase in the number of unresolved units. For example, respondents' mobility, unavailability and unwillingness have a negative effect on the possibility of making contact. In addition, progress in telecommunications technology has produced tools such as answering machines and call display that allow people to screen calls and avoid surveys. Another consequence of technological advances is that the same telephone numbers can be used for fax machines, Internet connections and residential land lines, a fact that compounds the difficulty of determining eligibility in some cases. Furthermore, due to the call limit, some cases that might have been resolved with a few additional calls are identified as unresolved.

For the RDD component of Cycle 21, since the target population was a subset of the general population, the category of unresolved units includes telephone numbers that reached a residence that did not satisfy the household composition matrix for one reason or another. The household composition matrix is a series of questions that collect data on the age, sex and relationship of household members. To be identified as an eligible unit, a household must be composed of at least one person aged 45 or over, and the composition matrix data are essential to confirm the household's eligibility. For example, a person reached at his or her private residence who refused to participate in GSS-21 before even completing the household composition matrix was an unresolved unit for GSS-21. Of the 57,741 telephone numbers sent to collection, about 18% were considered unresolved. Due to this high volume, a special methodology had to be developed.

5. Methodology used to calculate the response rate for the RDD component of GSS-21

5.1 Overview of the general methodology

For GSS-21, the case resolution methodology is similar to an imputation process in that a status is assigned at the micro level for each sampling unit deemed to be an unresolved unit. Imputation is carried out in two steps. First, a residence eligibility/ineligibility status is assigned; that is, it is determined whether or not the unresolved telephone number belongs to a private residence. Second, for all telephone numbers (both imputed and unimputed) that connect to a residence and have an incomplete or blank household composition matrix, an age eligibility/ineligibility status is assigned; this shows whether there is at least one person aged 45 or over in the household. The methodology has two parts: first, case eligibility is decided directly using a set of deterministic rules; second, an eligibility status is assigned on a random basis through modelling.

The main source of information used to resolve unresolved cases is paradata. Paradata are data from the collection process that assist not only in collection management but also in understanding and eventually optimizing the collection process (Laflamme, 2008). GSS paradata provide about 40 possible result codes for each call attempt: no contact, line busy, fax machine, refusal and so on. In addition, each telephone number sent to collection is associated with a vector of length 1 to 20. The vector contains the result code for each attempt made for that telephone number, along with the final collection status for the case: "final", "in progress" or "call limit reached".

5.2 Deterministic rules

At this stage, deterministic rules are used to determine case eligibility directly. The information sources used are paradata (possible result codes and case status) and survey data. The possible result codes were divided into three categories: (1) "residence" codes, which indicate that the telephone number is associated with a residence (e.g.

language barrier, refusal, appointment, etc.), (2) "non-residence" codes, which indicate that the telephone number connected to a unit that is not a residence (e.g. cell phone, collective dwelling, business), and (3) "indeterminate" codes, which indicate that no conclusion could be reached (e.g. no contact, answering machine, fax machine). With regard to the survey data, information collected at the beginning of the interview can help confirm whether there was contact with a private residence, and the data from the household composition matrix are essential to assigning an age eligibility. In constructing the rules, a hierarchical approach was taken in which priority was given to cases finalized in the field and result codes assigned by interviewers. The following tables show the rules for each level of eligibility to be determined.

Table 5.2-1
Summary of deterministic rules used for the RDD component of GSS-21 – Residence eligibility

For all cases:
1. Is the unit an eligible respondent?
➔ Yes: Residence eligibility = Yes
➔ No: Step 2
2. Was the unit assigned a final status in the field?
➔ Yes:
→ If the final result code is in the residence category: Residence eligibility = Yes
→ If the final result code is in the non-residence category: Residence eligibility = No
→ If the final result code is in the indeterminate category: Step 4
➔ No: Step 3
3. Is the unit a telephone number with unknown status that reached the five-call limit with no contacts?
➔ Yes: Residence eligibility = No
➔ No: Step 4
4. Analysis of the paradata vector: In the sequence of call attempts, is there a result code that belongs to the residence category?
➔ Yes: Residence eligibility = Yes
➔ No: Step 5
5. Analysis of the paradata vector: In the sequence of call attempts, is there a result code that belongs to the non-residence category?
➔ Yes: Residence eligibility = No
➔ No: Step 6
6. In the survey data, was it possible to confirm contact with a private residence?
➔ Yes: Residence eligibility = Yes
➔ No: Step 7
7. In the survey data, was it possible to confirm contact with an entity that is not a residence?
➔ Yes: Residence eligibility = No
➔ No: Step 8
8. Are more than half of the call sequence's result codes fax machine result codes?
➔ Yes: Residence eligibility = No
➔ No: Residence eligibility = Unknown

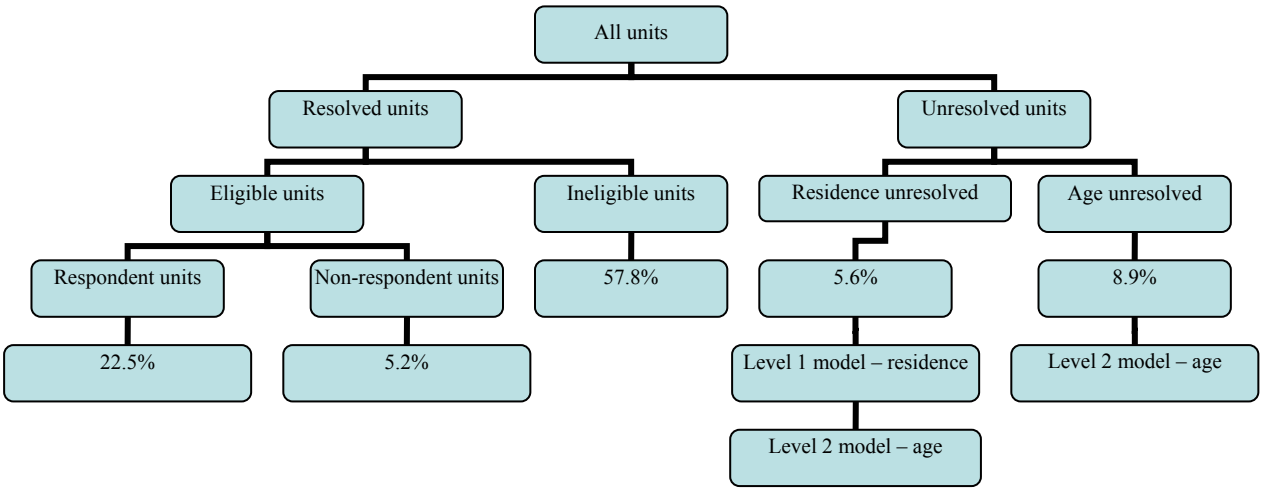
Table 5.2-2
Summary of deterministic rules used for the RRD component of GSS-21 – Age eligibility

For all cases where residence eligibility = Yes
1. In the survey data, has the relationship matrix been completed?
➔ Yes: Is there at least one person aged 45 or over in the household?
→ Yes: Age eligibility = Yes
→ No: Age eligibility = No
➔ No: Age eligibility = Unknown

Following the application of these deterministic rules, two categories of unresolved cases remain: (1) cases in which it is unknown whether the telephone number reached a residential household (residence eligibility = unknown in Table 5.2-1), and (2) cases in which the telephone number connected to a residence, but it cannot be determined

whether the household has at least one eligible member (age eligibility = unknown in Table 5.2-2). Figure 5.2-1 shows the breakdown of cases and the modelling stages. Out of all the units, 5.6% belong to the first category and 8.9% to the second category. As mentioned in section 5.1, modelling is carried out in two steps: (1) assignment of residence eligibility, and (2) determination of age eligibility.

Figure 5.2-1
Distribution of units in GSS-21 and modelling steps



5.3 Level 1 modelling – residence eligibility

For the level 1 model, a residence eligibility status is assigned randomly and independently for each unresolved telephone number in a homogeneity group (HG), with a particular probability. That probability is given by π_{HG}, where π_{HG} is the residence success rate (i.e. number of telephone numbers that reached a residence / number of resolved units) in the resolved portion of a GSS-21 HG.

The sources of information used to construct the HGs are paradata and the sampling plan. At this stage, the paradata information used is the total number of call attempts for the telephone number and collection wave information. The HG we selected is the result of intersecting the telephone number status (listed, unknown) with the stratum (urban, rural), the number of attempts (fewer than 10 calls, 10 calls or more) and the collection wave. In addition, each group must contain at least 20 units have a stable π_{HG} in each collection wave. When the model was implemented, 88% of the unresolved units were assigned a residence status. They are then combined with the age-unresolved units and assigned an age eligibility status via the level 2 model described in the next section.

5.4 Level 2 modelling – age eligibility

Step 1 provided a set of residential telephone numbers. Step 2 consists in determining an age eligibility status for those residential telephone numbers; that is, "What is the probability that a residential telephone line belongs to a household with at least one member aged 45 or over?" The procedure is the same as before: a status is assigned randomly and independently for each age-unresolved unit in an HG, with a particular probability θ_{HG}. For the age eligibility model, it was decided to use the strata as HGs (i.e. HG=STR). Four methods were considered for deriving θ_{STR}. Once this parameter has been computed and the model has been used, the denominator of the response rate shown in equation 1 can be calculated. The RDD component had 13,001 respondents.

Method 1
θ_{STR} is the age success rate (i.e. number of telephone numbers that reached an eligible household / number of telephone numbers that reached a residence) in the resolved portion of GSS-21 for each stratum. After this model was applied, the response rate was 61.1%.

Other methods

We also used another variant in which the proportion of residential telephone lines that belong to a household in the target population of the resolved and unresolved portion in each stratum is from another source. This allows us to derive another θ_{STR} that will be applied to the unresolved portion. In this case, θ_{STR} is given by

$$\theta_{STR} = \max\left(\frac{n_{(resolved+unresolved),STR}\,\theta_{other\ source,STR} - n_{resolved,STR}\,\theta_{resolved,STR}}{n_{unresolved,STR}}, 0\right)$$

We used the following sources:

Method 2: The outside source is the observed proportion in GSS-20. This allows us to work with a larger pool of telephone numbers, in which a slightly smaller proportion (7%) of unresolved cases was observed. The assumption in this case is that the observed proportion may be closer to reality. It is the proportion observed within the set of telephone numbers and is therefore unweighted. The response rate produced by this method is 61.1%, exactly the same as with method 1.

Method 3: The outside source is the proportion representing the historical average of the last three GSS cycles. The idea is to have an even larger pool of telephone numbers, but a less current pool, since it contains data going back to 2004. This is the proportion observed following the collection process, again unweighted. This method yields a response rate of 63.4%.

Method 4: The outside source is the proportion from the 2007 Residential Telephone Service Survey (RTSS). For each GSS stratum, we produced an estimate of the number of households that have at least one member aged 45 or over and can be reached via a residential telephone line out of all households that can be reached via a residential telephone line. This method produces a response rate of 59.5%.

The option selected was method 2. It yields the same result at the national level as option 1 and very similar results for each stratum. The same is true with option 4. When we calculate minimum and maximum response rates, based on the upper and lower limits of the RTSS estimate, respectively, we obtain 57.3% and 62%.

6. Response rate of the follow-up component and overall response rate

As described in section 2, the other component of GSS-21 is from a GSS-20 follow-up. In the case of a follow-up survey, the information about non-response must be based on the combined non-response of the primary survey and the secondary survey. The method used for this component of GSS-21 is simply the product of the GSS-20 response rate at the estimation stage and the GSS-21 response rate at the collection stage: 67.4% x 80.1% = 54%. An underlying assumption of this method is that the Cycle 20 response rate for the 45-and-over group is the same as for the under-45 group.

The overall response rate for the two components is given by

$$\text{Response rate} = \frac{\text{Respondent units}_{RDD} + \text{Respondent units}_{follow-up}}{(\text{Eligible units} + \text{Estimated eligible units})_{RDD} + (\text{Eligible units} + \text{Estimated eligible units})_{follow-up}}$$

It is important to note that in this equation, the number of eligible units and estimated eligible units from the GSS-20 follow-up component must be estimated. Since the target population for GSS-20 was persons aged 15 and over, some households may not have at least one person aged 45 or over. It was decided to simply divide the number of respondents in the GSS-21 follow-up component (10,403) by the component's response rate at the previous stage (54%). This yields an overall response rate of 57.7%. Another method would have been to model the Cycle 20 unresolved cases using an approach similar to the one described in section 5.

7. Conclusion

This paper illustrates another use of paradata. In addition to being essential for collection management, providing information and assisting in understanding non-response, paradata are pivotal for the calculation of the response rate, especially for an RDD survey. The response rate is an important measure of data quality, as it is the first step in determining the general magnitude of non-response in a survey. It is a complicated indicator to derive for an RDD survey, particularly an RDD survey that targets a subset of a population. Statistics Canada's guidelines provide a framework that must be adjusted to suit each survey's characteristics. GSS-21 is a good example because it shows the many alternatives that are possible at each stage of the non-response rate computation process. Choosing different deterministic rules, homogeneity groups and parameters in models, approaches and methods – all within the framework of the guidelines – has an impact on comparability between surveys. In some cases, particularly for GSS-21, the response rate is an estimated parameter. This demonstrates the limitations of using a single indicator to represent the quality of a complex survey.

References

Kish, L. (1976). Optima and Proxima in Linear Sample Designs, *Journal of the Royal Statistical Society A*, 139, 80-95.

Laflamme, F., Maydan, M. and Miller, A. (2008). Using Paradata to Actively Manage Data Collection Survey Process, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Statistics Canada (2000). Policy on Informing Users of Data Quality and Methodology. Policy Manual 2.3, Statistics Canada

Statistics Canada (2001). Standards and Guidelines for Reporting Nonresponse Rates. Statistics Canada technical report.