

# BATTLING ATTRITION IN THE NATIONAL LONGITUDINAL SURVEY OF CHILDREN AND YOUTH

Beatrice Baribeau<sup>1</sup>, Craig Wedseltoft<sup>2</sup> and Sarah Franklin<sup>3</sup>

## ABSTRACT

Non-response in a longitudinal survey has a cumulative effect because units are followed over time. This cumulative loss of sample is referred to as attrition. Attrition of the sample can lead to bias and reduced quality of estimates. The National Longitudinal Survey of Children and Youth (NLSCY) is a longitudinal survey having completed seven cycles of collection. The NLSCY will continue to survey those units sampled in cycle 1 indefinitely. Minimising attrition is a key challenge to the survey. The NLSCY also continually samples cohorts of young children for both longitudinal and cross-sectional analysis. For new cohorts, the effect of attrition on sample size can be anticipated at the sampling stage. This paper will discuss the efforts on fighting attrition in the NLSCY through maximisation of sample sizes for estimation and retention of respondents in the survey.

KEY WORDS: Attrition; Frame; Non-monotonic; Paradata

## RÉSUMÉ

La non-réponse dans les enquêtes longitudinales a un effet cumulatif puisque les unités sont suivies au fil du temps. Cette perte cumulative d'échantillons se nomme l'érosion de l'échantillon. Cette érosion de l'échantillon peut causer un biais et ainsi réduire la qualité des estimations. L'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ) est une enquête longitudinale qui a complété sept cycles de collecte. L'ELNEJ continuera d'enquêter sur ces unités échantillonnées au cycle 1 pendant une période indéterminée. La réduction de l'érosion de l'échantillon est un des principaux défis des enquêtes. L'ELNEJ échantillonne également de façon continue des cohortes de jeunes enfants tant pour les analyses longitudinales que pour les analyses transversales. Pour les nouvelles cohortes, les effets de l'érosion de l'échantillon sur la taille de l'échantillon peuvent être anticipés à l'étape de l'échantillonnage. Cet article analyse les efforts pour combattre l'érosion de l'échantillon dans l'ELNEJ au moyen de la maximisation des tailles d'échantillon pour l'estimation et le maintien des répondants dans l'enquête.

MOTS CLÉS : Base de sondage; érosion de l'échantillon; non-monotone; paradonnées

## 1. INTRODUCTION

### 1.1 Overview of the National Longitudinal Survey of Children and Youth

The National Longitudinal Survey of Children and Youth (NLSCY) is a longitudinal survey commenced in 1994. A new cycle of data is collected every second year. Collection of Cycle 7 was completed in the summer of 2007 and sampling of Cycle 8 will begin in the summer of 2008. A redesign of the survey is currently being undertaken.

Every cycle, the survey produces both longitudinal and cross-sectional information. The cohort of 0-11 year-olds initially sampled in 1994 is referred to as the original cohort. This cohort will be surveyed indefinitely and is representative longitudinally only. Every subsequent cycle has sampled a cohort of 0-1 year-old children and additionally 2-5 year-olds in some cycles. These cohorts are the Early Childhood Development (ECD) cohorts and are representative both longitudinally and cross-sectionally. The ECD cohorts are followed for a maximum of five cycles.

---

1 Beatrice Baribeau, Household Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6, [beatrice.baribeau@statcan.ca](mailto:beatrice.baribeau@statcan.ca)

2 Craig Wedseltoft, co-op student of Statistics Canada, University of British Columbia, Canada

3 Sarah Franklin, Household Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6, [sarah.franklin@statcan.ca](mailto:sarah.franklin@statcan.ca)

The NLSCY was designed to identify factors influencing the development of Canadians from birth to adulthood. The NLSCY collects detailed information about the factors influencing a child's cognitive, emotional and physical development and monitors the impact of these factors over time.

## **1.2 Attrition of the Sample**

Attrition is the loss of sample over time due to non-response. Maintaining a low rate of attrition is a key challenge in longitudinal surveys. Although it does not address bias, the effect of attrition on sample size can be countered for cohorts new to the NLSCY by increasing the initial size of the sample. However, the already determined original cohort has a non-increasing sample size requiring minimisation of the loss of respondents. This paper will discuss the efforts on fighting attrition in the NLSCY through maximisation of sample sizes given frame constraints and retention of the respondents already in the survey.

## **2. MAXIMISING SAMPLE SIZES**

### **2.1 Sample Design**

After Cycle 1, sampling has been restricted to the ECD cohorts, that is, no supplementary sample has been taken for the original cohort. In this paper, all efforts on maximising sample sizes refer to the ECD cohorts. Generally, the NLSCY sample consists of a subset of Labour Force Survey (LFS) responding households determined to be in-scope (i.e., having a child of the desired age). The LFS facilitates quick determination of in-scope households, is cost effective and allows for sampling of immigrant children. However, sampling from the LFS subjects the NLSCY to the constraints of LFS sample design, primarily a limited number of children.

The LFS is a nationally representative sample of 54,000 households. The primary focus of the LFS is to produce labour statistics and the survey therefore has a target population of 15-year-olds and older. Only 3% of Canadian households contain a baby born in 2006 and households containing children are not specifically targeted in LFS sampling and thus are not plentiful. Moreover, babies less than a year old are scarcer due to the way in which the NLSCY samples LFS households.

The NLSCY defines age by a reference year. The reference year for Cycle 7 was 2006, therefore in Cycle 7, all children born in 2006 are referred to as 0-year-olds, children born in 2005 are 1-year-olds, etc. The LFS is a panel survey, consisting of six panels referred to as rotation groups. Data is collected monthly. Each month a new rotation group is introduced and concurrently the oldest rotation group exits the survey. Rotation groups are surveyed for six consecutive months, creating 5/6ths overlap in the sample from month to month. The NLSCY primarily samples from the exiting rotation groups, although active rotation groups are sampled near the end of the NLSCY collection period in order to boost sample sizes. All household information is only as current as the latest LFS interview. For example, a rotation group that rotates out in March of the reference year will contain a list of all household members as of March. If a baby is born into the household in April of the reference year, the birth of that child will not be noted in the LFS file, thus the child will be ineligible for selection by the NLSCY. For this reason, only rotation groups rotating out in or after December of the reference year will provide complete coverage of births into households sampled in the LFS during the reference year. Collection of NLSCY data continues until June of the year following the reference year in part to increase the availability of rotation groups providing complete reference year coverage. Due to the rarity of households containing children born in the reference year, rotation groups not providing complete coverage are also sampled to boost overall sample sizes.

The Birth Registry, a database of all births by province that is maintained by Statistics Canada, was also used to sample 1-year-olds in Cycle 3 due to the plethora of children available on the registry. Historically, the LFS has been the preferred frame since it allows sampling of immigrant children and does not have the time delay experienced by the Birth Registry (approximately a year to collect all provincial information).

## **2.2 Sample Augmentation**

Increasing sample sizes for the longitudinal portion of the ECD cohorts can only be facilitated through an increase in the number of rotation groups used in the initial sample selection. However, the cross-sectional aspect of the ECD cohorts allow more options in increasing sample sizes. Some previous cycles have had demand for a large sample of 5-year-olds to measure readiness to learn (in school). To meet the analytical requirement, supplementary samples have been taken specifically for 5-year-olds in cycles 4 and 5, using the Birth Registry. More recently, in Cycles 6 and 7 a supplementary sample, or top-up, was taken from the LFS for 2-5 year-olds in all provinces other than Québec and Ontario. These top-ups increased the sample sizes so that cross-sectional estimates could be produced by province, not just region.

## **3. SAMPLE RETENTION**

### **3.1 A Non-monotonic Approach**

At the end of Cycle 6, the original cohort (with six cycles of collection/attrition) had a cumulative response rate of 62.1%. The 4-5 year-olds of the ECD cohort, with only three cycles of collection had a cumulative response rate of 60.1%. The original cohort, having completed three more cycles of collection than the 4-5 year-olds had lost fewer respondents due to a Swiss cheese, or non-monotonic approach. Due to the expected length of time original cohort members will remain in the survey, non-respondents are not automatically removed from the survey after an episode of non-response. Whereas, up until Cycle 7, for the ECD cohorts, a funnel or monotonic approach was applied, removing all non-respondents. For the original cohort, non-respondents are instead removed after two consecutive cycles of non-response. Longitudinally speaking, there may be gaps, or holes, in the response history of those surveyed. With the non-monotonic design, the attrition rate per cycle is on average 6.3%. The non-response per cycle is 13.6%, thus, on average, 7.3% of the sample is retained at each cycle by allowing the response history gaps.

### **3.2 Reintroduction of 18 year-olds**

The response unit for the NLSCY, that is, the interviewee, is the person most knowledgeable (PMK) about the child of interest. For children under 18 that are under the supervision of an adult, one of the adults, usually the mother, is determined to be the PMK. However, once the child has reached the age of 18, he/she becomes his/her own PMK. Youth that have been removed from the survey due to a history of non-response are now re-introduced to the NLSCY upon reaching the age of 18. The previous non-response was that of the previous PMK, not of the youth himself, so response histories are wiped clean. Re-introduction of youth previously dropped has yielded about a 33% response rate. Given the indefinite continuation of the survey for the original cohort, this has been considered a success and will be continued for future cycles.

### **3.3 Flagging Probable Non-respondents**

In Cycle 7 an innovative initiative was undertaken. Research on classifying non-respondents led to the creation of a binary variable which flagged co-operative and non-co-operative respondents based on the percentage of questions completed during the induction cycle. This flag was found to be a good predictor of future non-response. This variable was applied in the collection phase, with the intention that if households more probable to be non-respondents were known in advance, preventive measures such as assignment to experienced interviewers could reduce the likelihood of a non-response. In order to test the flag, a control group was implemented. Midway through collection, the control group was unmasked by the sampling methodologists to determine the effectiveness of the flag. The flag was found to have had no effect in most regions of the country; it was later determined that these regions had not used the flag for interviewer assignment or collection. The one region that implemented the flag experienced lower response rates with the flag than in the control group. This was later explained by the manner in which the flag was used. Interviewers were privy to which households were expected not to respond and thus put less effort (instead of more) into these households. Due to the sometimes ineffective and other times detrimental effect of the flag; it will not be used in future cycles.

### 3.4 Using Paradata to Concentrate Interviewer Efforts

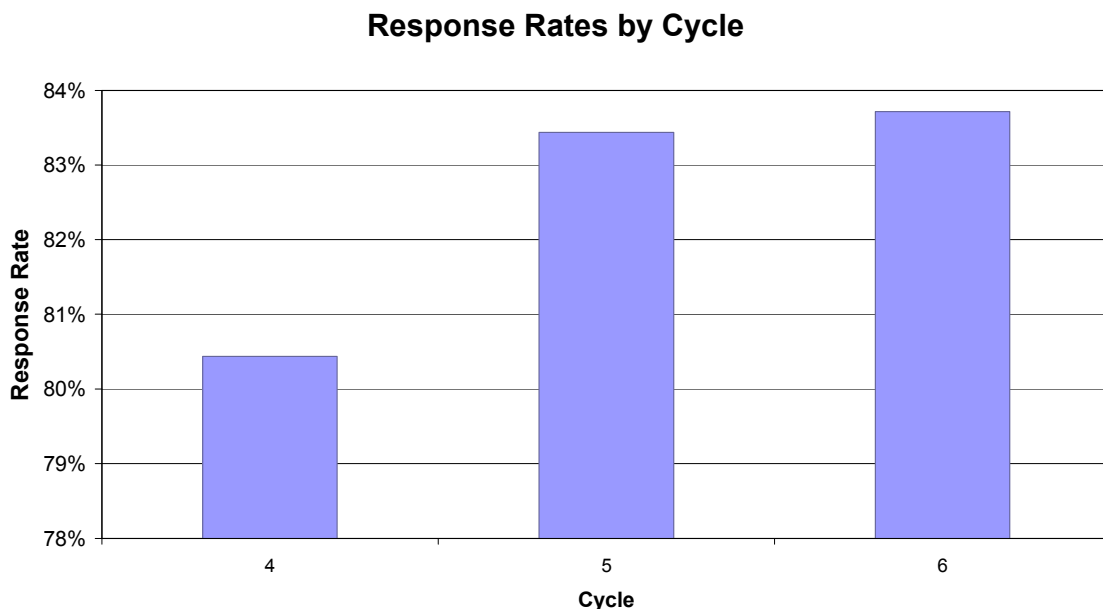
The paradata is the data stored during collection that contain information about collection such as the number of calls placed per household, the time of the call, the outcome of the call (not home, response, etc.). Analysis of this data creates a more complete picture of collection. Investigating this data can yield insights into the effectiveness of varying interviewer efforts in order to improve efficiency for future cycles.

The response rates for Cycles 3-6 of the NLSCY are shown in Graph 1. Cycles 5 and 6 are similar with an approximate 83.5% response rate while Cycle 4 lags slightly with an 80.5% response rate. By examining the differences between Cycle 4 and Cycles 5 and 6 we can learn what techniques were most effective in increasing the response rates. Three percent may seem a small difference, but in a longitudinal setting every respondent is extremely important due to the cumulative effect of non-response over time. Additionally, the target response rate of 88% was not achieved in any of the three cycles, so determining what techniques prove most effective could increase response rates beyond that of Cycle 5 or Cycle 6.

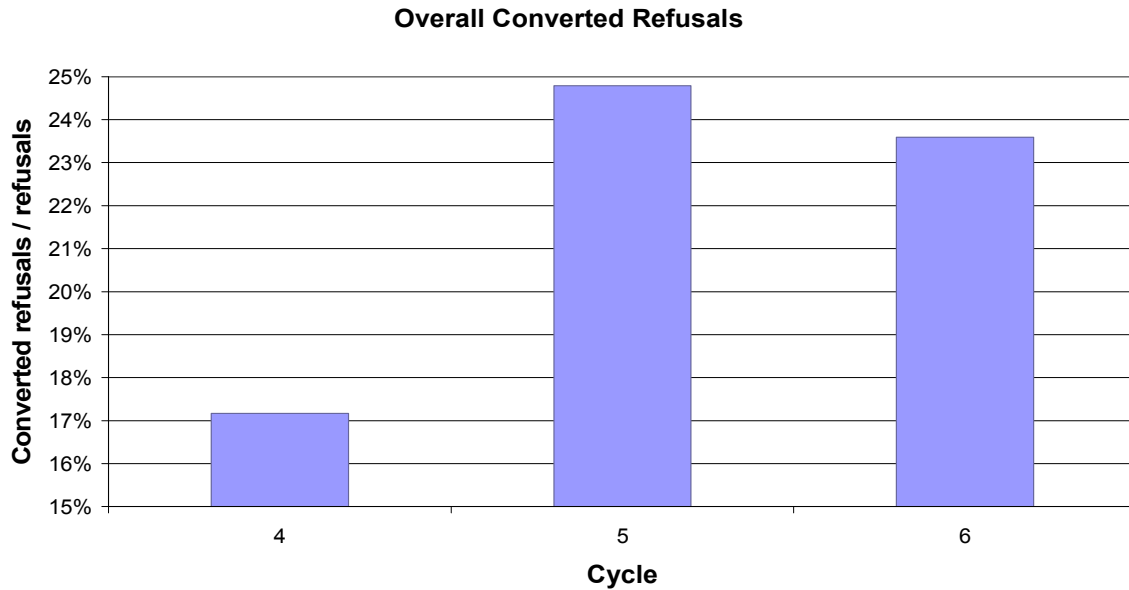
For the three cycles below, the largest source of non-response was refusals, contributing to over half of the non-response. A small percentage (less than 10% of non-respondents) was households that were not contacted, typically due to the inability to locate the respondent. The remaining non-response stems from a variety of sources such as absence during survey collection period, inability to obtain interview before end of collection period and language barriers.

The variable with the greatest discrepancy between Cycle 4 and Cycles 5 and 6 is the refusal conversion rate, that is, households that are converted from an initial status of refusal to a final status of response. The initial refusal rate is the percentage of households that at some point in the collection process refuse the survey. It differs from the final refusal rate in that some of the households that initially refuse are later converted to respondents. Interestingly, the initial refusal rate was similar across all three cycles whereas the final refusal rates differed. As seen in Graph 2, the conversion rate of refusing households in Cycle 4 is much lower than in Cycle 5 or Cycle 6. The effect of the lower refusal conversion rate on the response rate for Cycle 4 can be seen in Graph 3. The portion in black represents those households that would have been respondents had the conversion of refusals been in the range of Cycles 5 and 6. Note that 1/3 of the response rate discrepancy is now accounted for.

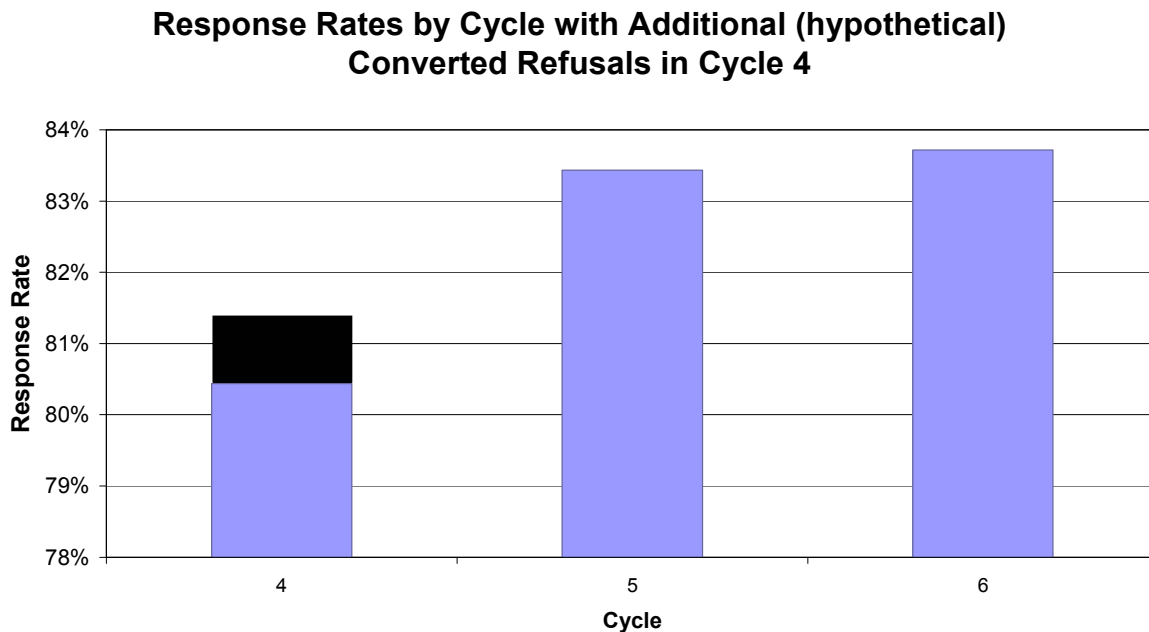
**Graph 1 – Response Rates for Cycle 4-6**



**Graph 2 – Conversion Rate for Refusing Households in Cycles 4-6**

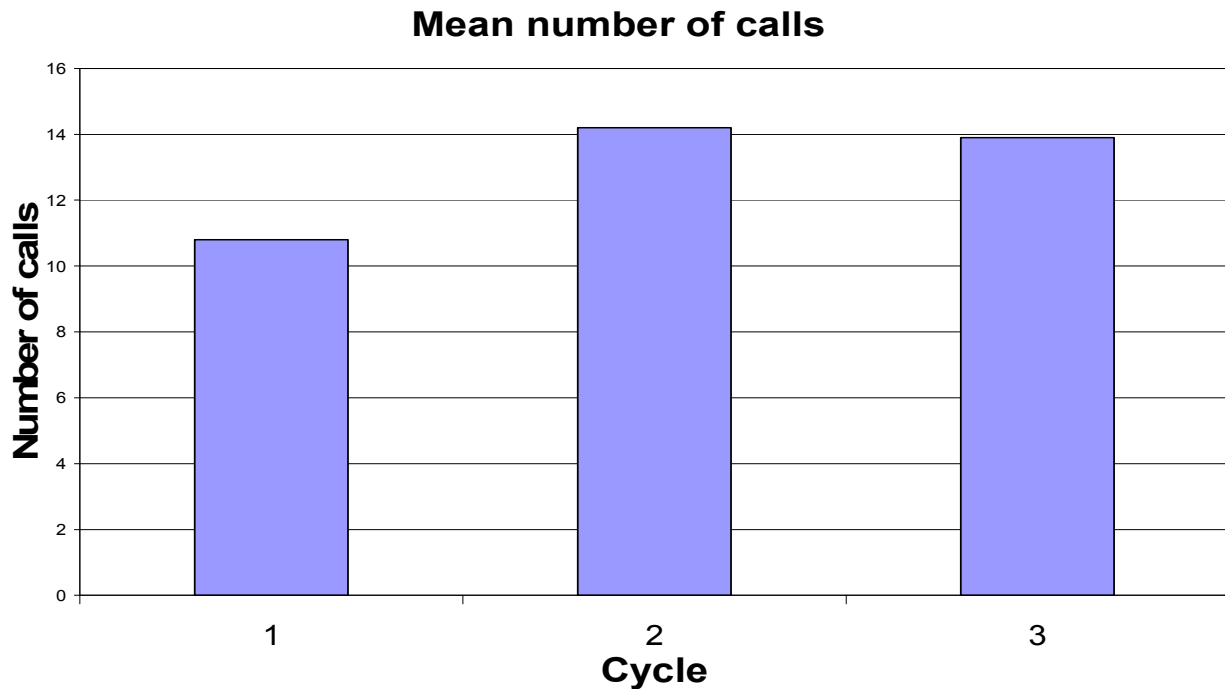


**Graph 3 – Response Rate Comparison Taking into Consideration an Increased Rate of Refusal Conversions**



The paradata further shows that the refusal conversion rate was lower in cycle 4 due to a lesser amount of effort put into calling refusing households. Graph 4 shows the mean number of calls placed to households that were still refusals at the end of collection. With an approximate 30% fewer calls placed to these households there were fewer opportunities to convert them to be respondents. Interviewer resources are limited but given that refusals contribute to over half of all non-response, that the biggest discrepancies between Cycle 4 and Cycles 5 and 6 are the refusal conversion rates and that equalising these differences (Graph 3) showed the largest impact on the response rates of all variables investigated, it is essential to allocate effort into calling households that have refused in order to convert them to responses.

**Graph 4 – Mean Number of Calls Placed to Households with a Final status of Refusal**



#### **4. FUTURE WORK**

The NLSCY is currently undergoing a redesign. The redesign will split the NLSCY into two separate vehicles: one cross-sectional and one longitudinal.

#### **ACKNOWLEDGEMENTS**

The authors would like to thank Cynthia Bocci, Wisner Jocelyn and Joanne Moloney for their valuable comments.

#### **REFERENCES**

- Bates, N. and Henley, M. (2006). "Using Call Records to Understand Response in Longitudinal Surveys". Paper presented at the Annual Conference of the American Association for Public Opinion Research, May 18-21, Montreal, Quebec.
- Bates, N. (2004). "Contact Histories: A Tool for Understanding Attrition in Panel Surveys". Paper presented at the Annual Conference of the American Association for Public Opinion Research, May 11-13, Phoenix, Arizona.
- Bates, N. (2003). "Contact Histories in Personal Visit Surveys: The Survey of Income and Program Participation (SIPP) Methods Panel". Paper presented at the Annual Conference of the American Association for Public Opinion Research, May 15-18, Nashville, Tennessee.