Improving the Unit Nonresponse Adjustment in the NLSCY Using Logistic Regression Modeling and Calibration

Mike Tam, Marcelle Tremblay, Sarah Franklin, Claude Girard
Statistics Canada, Main Building 2500 K, 150 Tunney's Pasture Driveway, Ottawa, ON K1A 0T6

1. Introduction

At its tenth anniversary, the National Longitudinal Survey of Children and Youth (NLSCY) underwent a thorough methodological review to evaluate the current state of the survey and opportunities for improvements. The results of this ten-year review were documented in Statistics Canada (2007). One of the findings of the report was that there was evidence of nonresponse bias in the estimates of the original cohort. In an effort to address this issue, some changes were made to the method of nonresponse adjustment. One was that the nonresponse model that was currently being used was replaced by a logistic regression model. Another was that instead of modeling incremental nonresponse from one cycle to another, a single model would be developed to model the cumulative nonresponse up to the current cycle. Finally, cooperation variables, which were indicators of item response propensity for a large selection of variables, were developed for use as predictors in the model. These changes were implemented for data collected at cycle 6 of the survey.

The ultimate goal of nonresponse adjustments is to minimize potential nonresponse bias. Unfortunately, since we do not have current information on the nonrespondents, it is not possible to apply a simple formula to estimate the amount of nonresponse bias such as, for example, the following formula for the bias of $\overline{Y}_r$ (Lessler and Kalsbeek, 1992, p.119):

$$\text{bias}(\overline{Y}_r) = \frac{nr}{n}\left(\overline{Y}_{nr} - \overline{Y}_r\right), \tag{1}$$

where $\overline{Y}_r$ and $\overline{Y}_{nr}$ are the sample means of the respondents and nonrespondents, respectively; and nr/n is the nonresponse rate. However, as described in Statistics Canada (2007), since we are concerned with a longitudinal survey, we can use the related concept of longitudinal consistency as a proxy for the bias due to nonresponse. (For a related discussion of *internal* consistency, see Singh *et al*. (1995)).

For the concept of longitudinal consistency that we use, suppose we estimate a cycle 1 characteristic with both the cycle 1 weights and the cycle t weights, where $t > 1$. Given that we are concerned with a longitudinal population, the difference between the two estimates should be due only to the additional nonresponse at cycle t. Hence, the nonresponse adjusted estimate at cycle t of a cycle 1 characteristic should be very close to the original cycle 1 estimate. We can also view this concept of longitudinal consistency as assuming that the cycle 1 estimate is our "best" estimate of a given characteristic since it has the lowest nonresponse. If we had confidence in the cycle 1 estimate, we should not gain confidence in a different estimate of the same characteristic that is based on less data. Therefore, we would like to minimize any deviation from the original cycle 1 estimate for cycle 1 characteristics.

In view of the preceding paragraph, to evaluate weights in terms of their ability to reduce nonresponse bias and maintain longitudinal consistency, we take the post-stratified weights at cycle t and compute the relative absolute differences between the actual cycle 1 estimates and the estimates of the cycle 1 characteristics computed using the cycle t weights. Since all of the estimates under consideration were of categorical variables, to ensure that categories with a small number of respondents did not distort the results, we calculated the absolute difference for each category of a variable and computed the total. This total was then computed as a percentage of the target population to arrive at the relative absolute difference. That is, if each variable k has J_k categories, we compute

$$r_k(t) = 100 \times \frac{\sum_{j=1}^{J_k}\left|\hat{Y}_{jk}(1) - \hat{Y}_{jk}(t)\right|}{\sum_{j=1}^{J_k}\hat{Y}_{jk}(1)}, \tag{2}$$

where $\hat{Y}_{jk}(1)$ and $\hat{Y}_{jk}(t)$ are the estimates for category j of variable k using the weights from cycle 1 and cycle t, respectively. The relative absolute difference (2) can also be seen to be the weighted average of the relative differences of the categories within a variable, with the weights being equal to the cycle 1 estimate of category j. Weights derived at cycle t that minimize

$r_k(t)$ for a large cross section of estimates would indicate that the cycle t weights are properly accounting for nonresponse.

In Statistics Canada (2007), a detailed investigation of the longitudinal consistency of several cycle 1 characteristics revealed that nonresponse bias was present for some variables. This led to the abandonment of the then current method of nonresponse weighting and the eventual selection of logistic modeling of cumulative nonresponse as the basis for weighting for nonresponse in future cycles of the NLSCY. The idea of longitudinal consistency also motivated the exploration of using calibration as a method to compensate for nonresponse. Using the information from cycle 1 as auxiliary data, calibrating the design weights of cycle 6 respondents to the original cycle 1 estimates was expected to yield cycle 6 weights which maintained longitudinal consistency.

In the remainder of the paper, we discuss the potential further improvements to the nonresponse model that had been contemplated, but which, due to time constraints, could not be put in place for cycle 6. In section 2, a brief overview of the NLSCY is given, including a description of the weighting methodology of previous cycles. Section 3 describes the additional developments to the cycle 6 logistic model for nonresponse and their impact on the consistency of estimates. Section 4 follows with a discussion of the feasibility of incorporating paradata (*i.e.*, information related to the collection or processing of survey data) into the nonresponse model. Given that our use of cooperation variables in the nonresponse model at cycle 6 proved to be very helpful in fitting the model, we felt that more detailed information such as that provided by paradata would also be of great use. Section 5 examines the use of calibration as an alternative to weighting class adjustments. Section 6 discusses the results of combining the logistic model with calibration. Since calibration cannot make use of paradata, this "hybrid" model would allow for both the incorporation of paradata and the use of calibration in the nonresponse adjustment. Finally, in section 7, we provide our conclusions and plans for future work.

2. Overview of the NLSCY

The NLSCY is a longitudinal survey, sponsored by Human Resources and Social Development Canada (HRSDC), which has been conducted by Statistics Canada since 1994. The objective of the survey is to collect information on characteristics and factors which impact on the development and well-being of Canadian children and youth over time. The survey samples households and data is collected every two years through telephone and personal interviews of the person most knowledgeable about the child (usually the mother), as well as direct measures such as math and vocabulary tests.

Data collection for the survey began in December 1994 with an original cohort of Canadian children aged 0 to 11 living in one of the ten provinces in Canada. Starting at cycle 2, Early Childhood Development (ECD) cohorts of 0- and 1-year-olds were selected and followed until they were 4 to 5 years old, respectively. In the fall of 2006, the survey began its seventh cycle of data collection.

The NLSCY produces longitudinal and cross-sectional weights for the ECD cohorts and two sets of longitudinal weights for the original cohort. The two sets of longitudinal weights adjust for the two types of nonresponse, monotone and non-monotone, respectively. In the monotone case, re-entries are discarded and units which are nonrespondents in a given cycle are deemed nonrespondents for all following cycles. In the non-monotone case, re-entries are allowed and so units may be nonrespondents in one cycle but respondents in a subsequent cycle. In this paper, we concentrate on the unit nonresponse of the original cohort at cycle 6 in the non-monotone case.

The following table gives the number of responding individuals in the original cohort in each cycle of the survey, as well as the percentage of the cycle 1 respondents who responded (non-monotone case) in each cycle.

Table 1. Number of respondents and the percentage of cycle 1 respondents still present at cycles 1 to 6

Cycle	Respondents (Child-level)	% of Cycle 1
1	16,903	100.0
2	15,468	91.5
3	14,997	88.7
4	13,310	78.7
5	12,523	74.1
6	11,483	67.9

Source: NLSCY Cycle 6 Microdata User Guide

Thus, by cycle 6, nearly a third of the cycle 1 respondents ceased responding to the NLSCY.

With the exception of cycle 3, for the first five cycles of the NLSCY, the nonresponse adjusted weight at cycle t was computed as the nonresponse adjusted weight at cycle $t-1$ adjusted for the nonresponse that occurred between cycles $t-1$ and t. At cycle 3, the nonresponse adjustment was applied to the cycle 1 nonresponse adjusted weight. The nonresponse adjustments were computed as the inverse of the response rate within weighting classes determined by the CHAID algorithm (Kass, 1980) and the nonresponse adjusted weights were then post-stratified to known demographic and geographic totals. As a consequence of the ten-year review of the NLSCY, it was decided that the nonresponse adjustment at cycle t be applied directly to the design weight. That is, a single nonresponse adjustment would be computed based on the cumulative nonresponse since cycle 1 up to, and including, cycle t, rather than computed sequentially as the product of adjustments between cycles $t-1$ and t, for $t>1$. A logistic regression model was used to model the cumulative nonresponse at cycle 6 and the resulting estimated probabilities of response, or propensity scores, were used to construct the weighting classes for the nonresponse adjustment. Nine response homogeous groups (RHGs) were used to adjust for nonresponse in the non-monotone attrition case. After adjusting for nonresponse, post-stratification based on age, sex, and province was performed. For further details on the weighting methodology of each cycle of the NLSCY, see Tremblay *et al*. (2007).

3. Logistic Modeling

The logistic models used at cycle 6 of the NLSCY were of the usual form

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}'\boldsymbol{\beta}\,, \qquad (3)$$

where π_i represents the cumulative response probability of the i^{th} individual to the current cycle. The model employed at cycle 6 incorporated frame variables and two cooperation variables. Survey data from cycles 1 to 5 were not used because they were not available for all units. The cooperation variables included in the model were derived as a way for item nonresponse at cycle 1 to be used as a predictor of subsequent unit nonresponse. The cooperation variables represented the proportion of selected cycle 1 survey questions which went unanswered by a given respondent. Two cooperation variables were computed, one in which questions to teachers and principals were included in the calculation and one in which those questions were excluded. After determining the values of each of the cooperation variables for each individual, the values were grouped into two sets of deciles. The grouped variables were both found to be significant in the cycle 6 model. (Tremblay *et al*. (2007) provides additional details.) There was a relatively small amount of unit nonresponse at cycle 1; for these units, the cooperation variables were imputed.

In order to incorporate further survey information into the response model at cycle 6, the problem of unit nonresponse at the previous cycles of the NLSCY first needed to be addressed. The additional difficulty here was due to the fact that we were dealing with non-monotone attrition and cumulative nonresponse. In the case of monotone attrition with sequential nonresponse, a response model could have been derived using the information from the previous cycle which would be available for all the respondents at the current cycle as well as for the nonrespondents who had responded at the previous cycle. Alternatively, for modeling cumulative nonresponse in a monotone attrition situation, the use of the previous cycles' data could be maximized by forming response models based on cycle response patterns as described in Lepkowski (1989). The last method could also have been used with non-monotone attrition; however, this would have resulted in too many patterns with very few units.

To incorporate as much of the previous cycles' data as possible, it was decided that imputation would be used. The use of imputation to handle cycle (wave) nonresponse has been mentioned in, for example, Lepkowski (1989) and Kalton (1986). To keep the process at a reasonably simple level, the imputation was done as follows. The original cohort was divided into a group of individuals who responded to all cycles and a group which had been nonresponding to at least one cycle. The first group was then considered as the pool of eligible donors. This was done to preserve the consistency of donors over the five cycles. For example, imputed data for nonrespondents at cycles 2 and 3 would come from the same donor. Propensity scores were then derived via logistic regression and cluster analysis on the propensity scores was used to form the imputation classes. Nearest neighbor imputation, based on a small selection of frame variables, was then done within the imputation classes.

Once complete datasets were derived using imputation, the variables for the response model had to be selected. The number of potential predictors presented by each cycle was large: approximately 1,800 variables were available at each cycle. Once the 100 or so variables from the frame were included, there were nearly 10,000 variables to consider for the model. To reduce this to a more tractable number, chi-square tests were used to first determine which variables were most related to response status. Then, in order to obtain as wide a cross section of variables as possible, for each questionnaire section, the variable with the smallest *p*-value and at most ten categories was selected. This was done for each cycle. Finally, adding the cooperation variables, approximately 180 potential covariates were determined and entered in a stepwise logistic regression procedure using the SAS system.

Several models were produced with varying numbers of predictors. A model with 54 predictors appeared to fit relatively well. The Hosmer-Lemeshow statistic was 0.156 and Nagelkerke's max-rescaled R-square (Nagelkerke, 1991) was 0.2. The area under the ROC (Receiver Operating Characteristic) curve, 0.726, also indicated a satisfactory fit. A model was also derived which minimized the number of imputed covariates by using only newly introduced variables. That is, for characteristics that were present in more than one cycle, only the variable representing that characteristic from the earliest cycle was kept. For example, if the characteristic *race* was introduced at cycle 2 and occurred again at cycle 3, only the variable representing *race* at cycle 2 would be kept. The rationale being that the first cycle in which a variable existed would have had the least nonresponse. The fit for this model was also satisfactory although inferior to the first model.

The consistency measure (2) was computed for the cycle 6 weights derived using only frame and cooperation variables and the cycle 6 weights based on the two new nonresponse models. The consistency measure was also computed for weights derived from the uniform nonresponse model for a baseline comparison. The uniform nonresponse model corresponds to the case where nonresponse is assumed to be random throughout the population rather than within subpopulations; that is, where $\pi_i = \pi$, for all $i \in U$, as opposed to $i \in U_h$, for some RHG h. The following table contains the usual summary statistics of the distribution of the consistency measure $r_k(6)$ when computed for variables $k = 1, \ldots, K$, where $K = 1{,}754$. The columns headed by *Q1*, *Med*, *Q3*, and *Max* indicate the first quartile, median, third quartile, and the maximum, respectively. Using a large number of variables and considering the distribution of the $r_k(6)$ provides us with a global view of how nonresponse is affecting the estimates. We can, of course, still evaluate specific variables to determine whether or not nonresponse bias is present and make the assumption that variables correlated with those selected behave similarly.

In the table, the following key is used:

- U - uniform nonresponse model;
- L1 - logistic model with only frame and cooperation variables;
- L2 - logistic model with frame and cooperation variables and variables from cycles 1 to 5;
- L3 - logistic model with frame and cooperation variables and only newly introduced variables from cycles 1 to 5.

Table 2. Summary statistics for the distribution of the $r_k(6)$ derived from different logistic models

Model	Mean	Q1	Med	Q3	Max
U	3.5	1.0	2.3	5.9	12.5
L1	1.6	0.8	1.4	2.3	8.7
L2	1.6	0.7	1.5	2.1	10.2
L3	1.5	0.7	1.4	2.1	9.5

From the table, the uniform nonresponse model clearly fails to maintain longitudinal consistency. The relative absolute differences of more than half the estimates derived under the assumption of uniform nonresponse exceeded 2%. In contrast, nearly 75% of the relative absolute differences of the estimates under L1 to L3 were under 2%. Amongst these three models, the differences between the distributions of the $r_k(6)$ are quite small. Though it did not have the best fit, its good performance with respect to longitudinal consistency leads us to favour model L3.

4. Incorporating Paradata into the Logistic Model

In addition to survey data, we also attempted to incorporate paradata (Couper and Lyberg, 2005) into the logistic model. The variables we considered were the number of attempts required to get the first contact, the number of attempts between the first contact and the final status, and whether tracing was required. We felt that together, these variables might help to determine a potential respondent's level of cooperativeness. For example, in the second case, whether an individual ultimately provided a fully or a partially completed questionnaire, the need for several follow-ups probably reflects a lack of cooperation.

Unfortunately, the only data for these variables available to the NLSCY were from cycles 4, 5, and 6. This presented a major difficulty since the NLSCY models nonresponse from the first cycle and therefore requires information from all cycles. However, nonrespondents from cycles 1 to 3 who have been dropped from the sample due to persistent nonresponse did not have any paradata defined at cycles 4 to 6.

Attempts at incorporating the paradata that was available were made such as deriving a variable indicating the missingness of the undefined paradata; however, this simply resulted in the paradata variables being proxies for prior nonresponse status. Under these circumstances, we felt that using the paradata would be unreliable.

5. Using Calibration to Improve Consistency

In view of our definition of longitudinal consistency, one way to ensure that the cycle 6 weights produce estimates of cycle 1 variables that are similar to estimates produced using cycle 1 weights is to use calibration estimation (Deville and Sarndal, 1992; Deville, Sarndal, and Sautory, 1993). See also Bethlehem (1988) and Lundstrom and Sarndal (1999) for the use of calibration as a method to address nonresponse.

In the usual calibration formulation, given initial weights (for example, design weights) d_i, the weights w_i are determined such that the distance between d_i and w_i, based on some metric $G(x)$, with $x = w_i / d_i$, is minimized and subject to the constraint, also called the calibration equations,

$$\sum_{i \in s} w_i \mathbf{x}_i = \mathbf{t}_x \ , \qquad (4)$$

where the $\mathbf{x}_i$ are auxiliary variables whose population totals, $\mathbf{t}_x$, are known. Deville and Sarndal (1992) provide several possible distances with which to implement calibration; however, we concentrated on two:

(i) The multiplicative or raking ratio method

$$G(x) = x \log x - x + 1 ; \qquad (5)$$

(ii) The logit (L, U) method

$$G(x) = \frac{1}{A}\left[(x-L)\log\frac{x-L}{1-L} + (U-x)\log\frac{U-x}{U-1}\right], \qquad (6)$$

if $L < x < U$, where L and U are two constants such that $L < 1 < U$ and $A = (U-L)/\{(1-L)(U-1)\}$.

The second distance could be used to set a lower bound on the w_i if any weight is deemed to have fallen below a minimum threshold.

The calibration was implemented using the software CALMAR 2 from INSEE (Sautory, 2003). Here, we use the actual cycle 1 estimates as auxiliary data. For given cycle 1 variables, estimates of the totals of the categories are determined. These totals then serve as the known marginals, $\mathbf{t}_x$, in the calibration. The initial weights, d_i, are the design weights. Once the calibrated weights have been determined, they are post-stratified to ensure that the distribution of the cross-classifications of age, sex, and province are reproduced.

The selection of variables for calibration was carried out using four methods. We initially requested from an analyst on the NLSCY a set of variables which were important from an analytical point of view. Calibrating on these variables would ensure that nonresponse bias was minimal for estimates involving these variables and that any analysis using these variables would not be adversely affected. The second approach to selecting the calibration variables was to use the significant cycle 1 variables from the chi-square tests described in Section 2. These variables were determined to be correlated with nonresponse; therefore, under the assumption that there is an underlying, unknown cause of nonresponse, correcting the estimates of these variables for nonresponse bias should also correct non-calibrated variables which were also related to nonresponse. The third method reduces the number of calibration variables by using the cycle 1 variables which were found to be significant in the logistic regression model L3 (see Table 2). In the latter two approaches, the cooperation variables were also included along with the cycle 1 variables. However, as the cooperation variables would not normally be estimated, it may make more sense to exclude them from the calibration. This was also done in a fourth calibration model. In all approaches, age, sex, and province were included as calibration variables if they were not already selected. This was done to reduce the effect of possibly large post-stratification adjustments of the weights after calibration.

Using the key:

C1 - variables important for analysis used for calibration (23 variables);

C2 - cycle 1 variables entered as potential predictors in logistic model used for calibration (41 variables);

C3 - cycle 1 variables considered significant by stepwise logistic regression used for calibration (16 variables);

C4 - same as C3, but excluding the cooperation variables (14 variables);

we have the following table containing the summary statistics of the $r_k(6)$ analogous to Table 2.

Table 3. Summary statistics for the distribution of the $r_k(6)$ derived from different calibration models

Model	Mean	Q1	Med	Q3	Max
C1	2.4	0.5	1.3	4.9	10.0
C2	1.0	0.4	0.8	1.3	8.2
C3	1.2	0.5	0.9	1.5	8.2
C4	1.3	0.5	1.0	2.0	8.2

The results in Table 3 include the values of $r_k(6)$, based on the same set of 1,754 variables as in Table 2, including the calibration variables. Although $r_k(6)$ for the calibration variables are highly likely to be zero (though not certain due to post-stratification), the small number of these variables did not cause them to affect the overall distribution. As can be seen in Table 3, the results for C2 and C3 are very similar. The relative absolute differences are quite small for 75% of the variables. The estimates produced under C1 do considerably worse; however, these variables were not expected to be related to nonresponse and they possibly were not correlated with a large number of variables. Given the much smaller number of variables required for virtually the same results, a calibration model with 16 variables (C3) appears to offer a large benefit in ensuring longitudinal consistency. In terms of specific calibration variables which were very useful, the cooperation variables were integral in reducing $r_k(6)$ for variables falling in the third quartile for model C3. Removing the cooperation variables from C3 to obtain C4 resulted in a similar distribution of $r_k(6)$, but with a larger third quartile of 2.0.

When we used the variables in C1 in the raking ratio method, some of the resultant weights were equal to zero. We tried to use the logit (L, U) method to produce non-zero weights; however, the algorithm failed to converge. These are two examples of the disadvantages associated with the calibration approach. Depending on the distance used, the constrained minimization may fail to converge. Even if the minimization is successful, the resultant weights may not be in an acceptable range. In addition, although we used paradata as auxiliary data in a logistic model to model the sample, paradata conceptually is not applicable as calibration variables since we would not normally estimate these quantities or use them to make inferences about the population.

6. Calibrating after Nonresponse Adjustment

Both the logistic regression and calibration models do a reasonably good job at adjusting for nonresponse, with calibration performing better under our measure of longitudinal consistency. On the other hand, variables such as paradata are more appropriately handled by the logistic models. Therefore, computing the calibration estimator after adjusting for nonresponse via logistic modeling and weighting within subgroups may provide the benefits of both methods.

In the NLSCY, we divide the responding sample into RHGs based on the propensity score of each respondent to provide some robustness against model misspecification. One of the benefits of the combined, or hybrid, approach is that we can add further protection against model misspecification by using variables such as the cooperation variables as covariates. In principle, the same variables are available for calibration, but this would lead to a large number of calibration equations. A second benefit is that the nonresponse-adjusted pre-calibrated weights can themselves be calibrated on different sets of totals depending on the analysis required. Thus, we can provide a set of general purpose weights reflecting our best efforts to adjust for nonresponse while at the same time offering some flexibility in their use.

Both calibration across all RHGs and calibration within RHGs were done; however, there were not major differences in the results. In what follows, only calibration across all RHGs is reported. The hybrid models that we looked at were:

L3/C3 - a logistic model containing frame and cooperation variables and variables that were newly introduced in a cycle; a calibration model where the variables included the cooperation variables;

L3/C4 - where the cooperation variables were not included in the set of calibration constraints.

Table 4. Summary statistics for the distribution of the $r_k(6)$ derived from different hybrid models

Model	Mean	Q1	Med	Q3	Max
L3/C3	1.2	0.5	0.9	1.5	9.1
L3/C4	1.2	0.5	0.9	1.5	9.1

To one decimal point, the consistency measures are the same for both hybrid models, so in the hybrid approach, omitting the cooperation variables from the calibration does not seem to be detrimental to the nonresponse adjustments. The hybrid results actually perform the best in terms of longitudinal consistency. This may be because, although the calibration model here (C4) does not include the cooperation variables, the logistic model prior to calibration in the hybrid model does.

7. Conclusion and Future Work

In this paper, we have discussed our experiences in trying to further improve the nonresponse model for the NLSCY in the case of non-monotone nonresponse for the original cohort of children. We have found that we are able to obtain good results for the combined RHG and calibration approach. This approach offers the opportunity for explicit modeling as well as ensuring consistency through calibrating to cycle 1 estimated totals. Future work may involve concentrating on specific variables or domains and extending our results to other cohorts of the survey.

Acknowledgements

We would like to thank our colleague Michel Ferland for developing and implementing the initial imputation strategy upon which the final imputations described in this paper were based.

References

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Couper, M. and Lyberg, L. (2005). The use of paradata in survey research. *Bulletin of the International Statistical Institute*, 55th Session, Sydney, Australia.

Deville, J.-C. and Sarndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Deville, J.-C., Sarndal,C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303-314.

Kass, G.V. (1980). An explanatory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 119-127.

Lepkowski, J.M. (1989). Treatment of wave nonresponse in panel surveys. In Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M.P., eds., *Panel Surveys*. New York: Wiley.

Lessler, J.T. and Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. New York: Wiley.

Lundstrom, S. and Sarndal, C.-E. (1999). Calibration as a standard treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.

Nagelkerke, N.J.D. (1991). A Note on a general definition of the coefficient of determination. *Biometrika*, 78, 3, 691-692.

Sautory, O. (2003). CALMAR 2: A new version of the CALMAR calibration adjustment program. *Proceedings of the Statistics Canada's Symposium 2003*.

Singh, A.C., Wu, S., and Boyer, R. (1995). Longitudinal survey nonresponse adjustment by weight calibration for estimation of gross flows. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 396-401.

Statistics Canada. (2007). *The National Longitudinal Survey of Children: 10-Year Review of the Methodology*. Sarah Franklin, ed. Household Survey Methods Division. Statistics Canada Internal Document.

Tremblay, M., Tam, M., Franklin, S., and Simard, M. (2007). Treating unit nonresponse in the Canadian National Longitudinal Survey of Children and Youth. *Submitted to the Journal of the Royal Statistical Society Series A*.