YOUTH IN TRANSITION SURVEY: A CASE STUDY IN DESIGN AND DEVELOPMENT

Joanne C. Moloney and Mike A. Hidiroglou, Statistics Canada
Joanne C. Moloney, Statistics Canada, Ottawa, Ontario Canada K1A 0T6, joanne.moloney@statcan.ca

Key Words: Dual-frame design; under-coverage bias; longitudinal surveys; misclassification error; non-sampling error

1. Introduction

The Youth in Transition Survey (YITS) is a new Statistics Canada survey designed to collect longitudinal data on school-to-work transitions for two age-specific cohorts, 15-year-olds and youth aged 18 to 20. The YITS project was initiated in 1996. At that time the Canadian government identified school-to-work transitions as one of the key areas where more information was needed for policy development.

YITS has been designed to identify and collect information on "at-risk" subgroups of the population. High-school leavers, that is, persons who have not received a high-school diploma or certificate and are not attending high school, comprise a domain of primary interest in studying school-to-work transitions.

Understanding these processes requires longitudinal data and ideally, the data collection should begin with persons below the legal age for leaving school. For this reason, a cohort of 15-year-olds is included in YITS and will be surveyed every two years for a planned total of five cycles. The school-based sample design for this cohort accommodates skills testing of students and measurement of school effects.

Information from previous surveys suggests the proportion of high-school leavers may peak among 20-year-olds. To provide more immediate data on this group of youth, a cohort of 18-20 year-olds is included in YITS, with the intention of surveying them every second year for up to three cycles. The sample design of the 18-20 cohort is the subject of this paper.

The YITS target population for the 18-20 cohort comprises residents of the ten provinces of Canada who were born in the years 1979 to 1981. A large portion of the questionnaire for Cycle 1 (conducted from January to April, 2000) relates to education and labour market activities during the reference year of 1999, when persons born in 1979 to 1981 would be 18 to 20 years old.

In Canada education is a provincial responsibility and there are important differences among the educational systems in place. There is clearly a need for certain population estimates at the province level, for example, leaver rates (or proportions of other at-risk groups) among 20-year-olds by sex at Cycle 1.

This paper describes the dilemma of choosing a design that would provide a suitable sample for the highly mobile and relatively rare 18-20 cohort. The approaches considered for a single-frame design, as described in Section 2, eventually led to the suggestion of a dual-frame design. The rationale and theory of the dual-frame sample design using a screening estimator and a full-dual estimator for YITS are presented in Section 3. Section 4 discusses the sample sizes computed for the dual-frame estimators and compares them to sample sizes required for an alternative single-frame design. The results of the pilot survey are outlined in Section 5. These findings led to the adoption of a single-frame design for YITS, presented in Section 6. Section 7 provides concluding remarks.

2. Approaches considered for a single-frame sample design

The need for estimates of small domains within each province over a series of three cycles posed a dilemma for the YITS sample design. With these requirements it was estimated that if the sample were selected from an up-to-date household frame, the initial sample size would be about 20,000 persons. A sample selected from an older frame would have to be inflated to take into consideration the mobility rates of 18-20 year-olds. Although the rates vary considerably by sex and by province, 1996 Census data indicate that for the age group as a whole, 22% were living in a different dwelling in May 1996 from the dwelling they occupied one year earlier.

Several options were examined in the search for a sample design based on a single frame. As a predecessor of YITS, first we examined the frame used for the Statistics Canada School Leavers Survey (SLS) conducted in 1991. This survey had the same target population as YITS and similar but less extensive content. Recent and current administrative files on parents and children receiving benefits from the Family Allowance program were used to create the SLS frame. This federal government program at that time provided benefits to parents of any child up to the age of 18, provided the child was still attending school. The administrative files identified cases in which benefits were terminated due to the child leaving school. Unfortunately for YITS, the Child Tax Credit, a means-based instrument that could provide only a partial frame

for all 18-20 year-olds, replaced the Family Allowance program.

We also considered a random-digit dialling (RDD) design. However, fewer than 10% of households have a member in the 18-20 age group and of those that do, only about 10% have more than one. Based on an average "hit rate" of 1 in 10, and an average of two calls per contact, the RDD option implied a very high cost per completed interview.

No assessment of household frames would have been complete without considering the Labour Force Survey (LFS). The LFS is a monthly survey that collects labour market data from a sample of 60,000 dwellings, comprised of six rotation groups. [1] The sample of dwellings is selected according to a multi-stage stratified clustered design. (See Gambino, Singh, Dufour, Kennedy and Lindeyer, 1998). Dwellings are in the LFS for six consecutive months. Each month the dwellings in one of the rotation groups are replaced by a new sample. The first month a dwelling is in the LFS a roster containing information on the household composition is completed. The roster includes the name, sex, date of birth and education level of every household member. When the dwelling is contacted in subsequent months for the LFS the roster is updated to reflect changes in household membership from the previous month.

Traditionally the LFS has provided the sample for many cross-sectional surveys, some of which are conducted as supplementary questions at the end of the LFS interview. In recent years longitudinal surveys have begun to use the current LFS sample, dwellings that have recently rotated out of the LFS, or freshly listed households from the LFS frame. To minimise the cost for a sample with up-to-date household composition and contact information, the optimal choice would be to use the set of dwellings currently in the LFS. However, preliminary estimates indicated many more than six rotation groups would be required to obtain reliable survey estimates for YITS. Obtaining freshly listed households from the LFS frame was considered too costly, given the low incidence of dwellings with 18-20 year-olds. This left the option of using the current LFS sample augmented by households that had rotated out of the LFS in previous months. Although this would provide a cheap frame, the idea of trying to trace and contact adolescents whose households had been in the LFS one or two years in the past was daunting. Given the experience of other surveys based on LFS rotate-out groups, tracing these individuals would be a challenge. Furthermore, from the perspective of the LFS capacity to support other potential household surveys, the option of YITS using such a large number of rotation groups to cover such a narrow segment of the population seemed inefficient. Despite these drawbacks however, the LFS with current and rotate-out dwellings was not entirely ruled out.

Attention turned briefly to administrative sources such as electoral lists and tax files. However, for both these sources there were issues concerning the coverage of the target population and record linkages to obtain telephone numbers.

Finally, from the perspective of the sample size and coverage needed, the Canadian Census of Population seemed a suitable option to consider for YITS. The Canadian Census is a quinquennial household survey. It uses two principal questionnaires for data collection. The short questionnaire (Form 2A) collects basic information such as the date of birth and sex of each household member and their relationship to one another. The longer questionnaire (Form 2B) includes all the Form 2A content plus a host of other questions relating to ethnicity, education, mobility, labour market activity and income. About 4 out of 5 dwellings in Canada receive the Form 2A; the remainder receive the Form 2B.

In addition to the coverage of the Census frame and the implied available sample size, it seemed design efficiencies might be achieved by restricting the YITS sample to the 2B households. For example, data on the education and income levels of all household members could be employed to try to target adolescents potentially "at-risk".

But the Census frame had its drawbacks as well. First, waiting for the 2001 Census data was not an option, given the YITS timeframe, but the 1996 Census data would be more than three years old by the time of the YITS Cycle 1 data collection. Over a time period this long many of the adolescents in the target population would have moved to other dwellings. Moreover, using the 1996 Census as a frame of individuals to be sampled and traced by name was not permitted. The Census could be used only as a frame of dwellings and even this process would require special permission based on a lack of alternatives. In the case of YITS, the frame would be used to identify dwellings that, at the time of the Census, had a household member in the target population.

It was feared this approach would lead to biased survey estimates for YITS, given that characteristics such as mobility might not be independent of characteristics that define youth "at-risk", such as

[1] LFS coverage of the population 15 years of age and over excludes full-time members of the Canadian Armed Forces, inmates of institutions, residents of Nunavut, the Northwest Territories and persons living on Indian reserves – together these groups represent about 2 % of the population 15 years of age and over.

leaving high school, for example. Estimates of mobility rates among 18-20 year-olds computed from 1996 Census data indicate that mobility and being a high-school leaver are not independent (Table 2.1).[2] Not surprisingly, for the 20-year-old population the differences are even more pronounced. Over the period of one year 20-year-old leavers are about 65 percent more likely than non-leaver 20-year-olds to change their usual place of residence.

Table 2.1 Estimated mobility rates by age and leaver status

Age as of Census day (years)	Time period (years)	Leavers	Non-leavers
18-20	1	0.36	0.19
	5	0.58	0.39
20	1	0.40	0.24
	5	0.63	0.42

Despite its disadvantages, the use of the Census was favoured among all the options that were considered. However, the concern over potential bias due to under-coverage of movers led to consideration of a dual-frame design for YITS.

3. Dual-frame design

The application of multiple-frame methodology for the YITS sample design entails a sample selection from two frames, that is, the LFS frame and the 1996 Census of Population frame. This seems a rational approach to compensate for the age of the Census frame: essentially the LFS is used to cover the part of the target population that would be missed if the Census frame alone were used to select the YITS sample.

To implement the dual-frame option, the YITS target population is partitioned into two domains, movers and non-movers. Movers are persons whose usual place of residence at the time of YITS is different than that at the time of the Census, that is, on May 14 1996. Technically the mover domain also includes immigrants and households in dwellings constructed between the Census and YITS. Non-movers comprise the rest of the population – note that this domain includes individuals who move out of their Census dwelling and then move back to the same dwelling and live there at the time of the YITS data collection.

As previously noted, the names of individuals are not available on the 1996 Census frame. Therefore, the sample units selected from each frame correspond to a set of in-scope dwellings. These are dwellings that, according to the frame, contain a household member born in the years 1979 to 1981. Of course, by the time these dwellings are contacted for YITS, the household membership will have changed for some of the dwellings, especially those in the Census sample. Therefore, on contacting each sample dwelling it is necessary to determine how many current household members are in the YITS target population, and for each, whether he/she is a mover or non-mover.

To reduce respondent burden at the household level, only one respondent per dwelling can be selected for YITS. The simplest strategy to cover the target population through the two samples is to interview non-movers in the Census sample and movers in the LFS sample. Alternatively, it might be preferable to also interview non-movers in the LFS sample rather than "wasting" these sample units once the dwelling is contacted. The two possible options for determining whom to interview in each sample are illustrated in Table 3.1.

Table 3.1 Interview status for dwellings within each frame

Household composition at YITS		Who to interview	
Movers	Non-movers	Census sample	LFS sample
No	No		
Yes	No		Mover
No	Yes	Non-mover	Non-mover?
Yes	Yes	Non-mover	Mover

The dual-frame approach offered a possible solution for the YITS sample design. However, this option relied on the classification of respondents as movers or non-movers. There were doubts concerning the ability of respondents to provide accurate information to determine their mover status, especially considering the transient age group targeted by YITS. To help determine whether the dual-frame approach would be feasible for YITS, testing the reliability of mover status became one of the objectives of the pilot survey. (See Section 5 for details.)

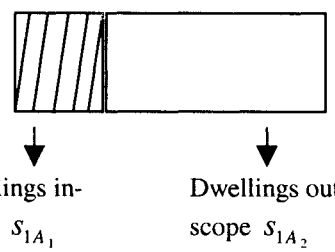
[2] For an out-of-date frame such as the Census, a positive correlation between mobility and becoming a high school leaver would be a counter-argument for trying to target youth in "at-risk" groups through the stratification variables in the sample design.

At the same time, more information was required concerning the estimators and required sample sizes for the dual-frame, assuming the mover status of respondents was known. The remainder of this section discusses the technical aspects of the design and develops the dual-frame estimators.

3.1 1996 Census 2B sub-sample

The Census 2B sample is a one in five systematic sample of dwellings drawn at the Census enumeration area level.[3] For purposes of notation simplification, we do not include stratification in the notation related to the Census estimates. The total number of dwellings in Canada, comprising the universe U_A, is denoted as N_A. The Census 2B sample s_{1A} contains $n_{1A} \doteq 0.2\, N_A$ dwellings and is split into two parts: dwellings in-scope and out-of-scope for YITS. A dwelling is in scope for YITS only if any of its members were born in 1979 to 1981. The in-scope sample s_{1A_1} has n_{1A_1} dwellings, whereas the out-of-scope sample s_{1A_2} has n_{1A_2} dwellings, as depicted by Figure 3.1.

Figure 3.1 Status of Census 2B YITS dwellings



The first-phase in-scope sample s_{1A_1} is stratified by province and household characteristics such that stratum $s_{1A_{1h}}$ ($h=1,\ldots, H$) has $n_{1A_{1h}}$ units. A sample $s_{2A_{1h}}$ of $n_{2A_{1h}}$ dwellings is then selected within each stratum $s_{1A_{1h}}$. The resulting second-phase sample $s_{2A_1} = \bigcup_{h=1}^{H} s_{2A_{1h}}$ has $n_{2A_1} = \sum_{h=1}^{H} n_{2A_{1h}}$ dwellings. Selected dwellings are clusters of individuals, and only one in-scope person (that is, a member born in 1979 to 1981) is selected within each one. Selected dwellings are interviewed to determine the mover status (mover or non-mover) of its in-scope residents. Note that the mover/non-mover status is a domain because it is not known before the sample is drawn. On contacting a sampled dwelling for YITS, the interviewer must determine the mover status of current household members born between 1979 and 1981. Sampled dwellings that do not have any members born in these years at the time they are contacted for YITS are ineligible for the survey. The sample s_{2A_1} is therefore further split into a domain of movers $s_{2A_1,M}$ with $n_{2A_1,M}$ dwellings, a domain of non-movers $s_{2A_1,\bar{M}}$ with $n_{2A_1,\bar{M}}$ dwellings and a third domain comprising $n_{2A_1,O}$ dwellings that are no longer eligible for YITS.[4] Corresponding domains for mover status at the stratum level are $s_{2A_{1h},M}$ and $s_{2A_{1h},\bar{M}}$. If there is at least one non-mover within the dwelling, then one of them is randomly selected and interviewed. Movers are not selected and therefore are not interviewed. Estimates of interest are computed for domains d that are associated with non-movers.

An estimate from the Census sample for the total of a given characteristic y in domain d is:

$$\hat{Y}_{A_1}(d) = \frac{N_A}{n_{1A}} \sum_{h=1}^{H} \frac{n_{1A_{1h}}}{n_{2A_{1h}}} \sum_{s_{2A_{1h}}} M_{2A_{1hi}}\, \bar{y}_{hi}(d)$$

where $M_{2A_{1hi}}$ is the number of individuals that are non-movers in the YITS target population in the i-th sampled dwelling in stratum h; $\bar{y}_{hi}(d)$ is the sample mean in the i-th sampled dwelling within stratum h. The domain population variance is given by:

$$V\left(\hat{Y}_{A_1}(d)\right) = N_A^2 \frac{1-f_{1A}}{n_{1A}} S_A^2(d) + \sum_{h=1}^{H} \frac{N_A \cdot N_{A_{1h}}}{n_{1A}} \left(\frac{n_{1A_{1h}}}{n_{2A_{1h}}} - 1\right) S_{A_{1h}}^2(d) + \sum_{h=1}^{H} \frac{N_A}{n_{1A}} \frac{n_{1A_{1h}}}{n_{2A_{1h}}} \sum_{i=1}^{N_{A_{1h}}} M_{2A_{1hi}}^2 \left(1-f_{2A_{1hi}}\right) S_{2A_{1hi}}^2(d)$$

where $f_{1A} = n_{1A} / N_A$ is the first-phase sampling fraction; $f_{2A_{1hi}} = m_{2A_{1hi}} / M_{2A_{1hi}}$ is the sampling fraction of individuals in the i-th dwelling within the h-th stratum; $S_A^2(d)$ and $S_{A_{1h}}^2(d)$ denote respectively the between-dwelling population variance of the characteristic y in the domain d in U_A and $U_{A_{1h}}$ (the portion of the universe U_A that contains dwellings in scope for the h-th stratum defined at the first-phase);

[3] An enumeration area is the geographic area canvassed by one Census representative.

[4] A dwelling in the sample from the Census frame is in the non-mover domain if it has at least one non-mover born in 1979 to 1981. A dwelling with no non-movers and at least one mover born in these years is in the mover domain.

and $S^2_{2A_{hi}}(d)$ is the within-dwelling variance for individuals in the i-th dwelling within the h-th stratum.

3.2 Labour Force Survey Sample

Let the sample size required by YITS from the LFS be n_B dwellings. The sample s_B is stratified into G strata, s_{B_g} (say), with each stratum consisting of n_{B_g} dwellings. Note that the overall sample size is $n_B = \sum_{g=1}^{G} n_{B_g}$. We denote the dwelling size within the i-th dwelling in stratum g as $M_{B_{gi}}$: a sample of $m_{B_{gi}}$ individuals is selected from this dwelling. The estimator of the total for a domain d and variable y is:

$$\hat{Y}_B(d) = \sum_{g=1}^{G} \sum_{i \in s_{B_g}} w_{B_{gi}} \hat{Y}_{B_{gi}}(d)$$

where $w_{B_{gi}}$ is the LFS design weight (adjusted for the number of rotation groups included in the YITS sample) for the i-th dwelling belonging to stratum g; $\hat{Y}_{B_{gi}}(d) = \sum_{k \in s_{B_{gi}}} \frac{M_{B_{gi}}}{m_{B_{gi}}} y_{gik}(d)$ is the estimate of the y-characteristic within the i-th dwelling within stratum g.

The population variance for the domain estimator $\hat{Y}_B(d)$ is given by

$$V\left(\hat{Y}_B(d)\right) = \sum_{g=1}^{G} \sum_{i=1}^{N_{B_g}} \frac{1}{w_{B_{gi}}} \left(w_{B_{gi}} Y_{B_{gi}}(d) - \frac{Y_B(d)}{n_{B_g}} \right)^2 + \sum_{g=1}^{G} \sum_{i=1}^{N_{B_g}} \frac{M^2_{B_{gi}}}{m_{B_{gi}}} w_{B_{gi}} \left(1 - \frac{m_{B_{gi}}}{M_{B_{gi}}}\right) S^2_{2B_{gi}}(d)$$

where, for the gi-th dwelling $Y_{B_{gi}}(d)$ denotes the true total and $S^2_{2B_{gi}}(d)$ is the within-dwelling sampling variance. The true population total is $Y_B(d)$.

The sample s_B is split into two parts: (i) s_{B_1} dwellings that are in-scope to YITS, and (ii) s_{B_2} dwellings that are not in-scope to YITS. Only the dwellings within s_{B_1} are interviewed. Let n_{B_1} and n_{B_2} be the corresponding samples sizes. As was the case for the Census 2B second-phase sample, the sample s_{B_1} is further split into the domains of movers $s_{B_1,M}$, non-movers $s_{B_1,\bar{M}}$ and a domain of dwellings no longer eligible for YITS at the time of contact.[5] Hence n_{B_1} is further split into $n_{B_1,M}$, $n_{B_1,\bar{M}}$ and $n_{B_1,O}$; $n_{B_1,M}$ is the observed sample size of movers, $n_{B_1,\bar{M}}$ is the observed sample size of non-movers and $n_{B_1,O}$ dwellings are no longer eligible for YITS.

Only the movers in the sample $s_{B_1,M}$ must be interviewed. However, collecting data for the non-movers $s_{B_1,\bar{M}}$ as well is another option. These data would have to be combined with the data obtained from the Census sample, using multiple-frame methodology. The question is whether the resulting gains in precision would be worth the complication in the methodology of YITS. If this multiple-frame design were implemented, the actual methodology for combining the data would most likely use the Skinner-Rao (1996) procedure as it yields unique factors that can be applied to all variables of interest.

3.3. Multiple Frame Procedure

3.3.1 Screening Estimator

The partial use of multiple-frame methodology (i.e. the screening estimator) implies that we use only the data from the movers in LFS sample and the data from the non-movers in the sample from the Census frame. That is, the estimator of the total for a given characteristic y and domain d becomes:

$$\hat{Y}_{SCREEN}(d) = \hat{Y}_{A_{1\bar{M}}}(d) + \hat{Y}_{B_{1M}}(d) \qquad (3.1)$$

where $\hat{Y}_{B_{1M}}(d)$ only contains information from the sample s_{B_1} of movers.

3.3.2 Multiple Frame Estimator

The full use of multiple-frame methodology implies that the estimator of the total becomes:

$$\hat{Y}_{MULT}(d) = \lambda \hat{Y}_{A_{1\bar{M}}}(d) + (1-\lambda)\hat{Y}_{B_{1\bar{M}}}(d) + \hat{Y}_{B_{1M}}(d) \qquad (3.2)$$

The factor λ is obtained by minimizing the variance of $\hat{Y}_{MULT}(d)$, given by:

$$V\left(\hat{Y}_{MULT}(d)\right) = \lambda^2 V\left(\hat{Y}_{A_{1\bar{M}}}(d)\right) + (1-\lambda)^2 V\left(\hat{Y}_{B_{1\bar{M}}}(d)\right) + V\left(\hat{Y}_{B_{1M}}(d)\right) + 2(1-\lambda) Cov\left(\hat{Y}_{B_{1\bar{M}}}(d), \hat{Y}_{B_{1M}}(d)\right)$$

[5] A dwelling in the sample from the LFS frame is in the mover domain if it has at least one mover born in 1979 to 1981. A dwelling with no movers and at least one non-mover born in these years is in the non-mover domain.

Differentiating with respect to λ and setting the derivative to zero we obtain:

$$\lambda V\left(\hat{Y}_{A_{1\bar{M}}}(d)\right)-(1-\lambda)V\left(\hat{Y}_{B_{1\bar{M}}}(d)\right)-Cov\left(\hat{Y}_{B_{1\bar{M}}}(d),\hat{Y}_{B_{1M}}(d)\right)=0$$

$$\lambda_{opt}=\frac{Cov\left(\hat{Y}_{B_{1\bar{M}}}(d),\hat{Y}_{B_{1M}}(d)\right)+V\left(\hat{Y}_{B_{1\bar{M}}}(d)\right)}{V\left(\hat{Y}_{A_{1\bar{M}}}(d)\right)+V\left(\hat{Y}_{B_{1\bar{M}}}(d)\right)} \qquad (3.3)$$

where λ_{opt} applies to the domain of interest d.

4. Computing sample sizes

This section describes how the sample sizes required to obtain reliable leaver rates were computed for alternative sample designs: i) the dual-frame design with a screened sample; ii) the dual-frame design with the full-dual estimator; iii) the LFS frame with a sample of individuals selected from active and rotate-out groups.

Assuming the mover status could be correctly identified, there were other questions to answer before a decision could be made regarding the YITS sample design. What level of data quality could be obtained using a dual-frame design? Would it be preferable to use partial or complete multiple-frame methodology? How would the quality of estimates obtainable from the dual-frame compare to those resulting from the single-frame design?

As previously noted, one of the important types of estimates to be provided by YITS is the Cycle 1 leaver rate among 20-year-olds, by sex and province. For the purpose of estimating the required sample size for these leaver rates, we consider the corresponding set of target populations separately, each comprising only persons born in 1979.

4.1 Frame vintage versus mobility

The two dual-frame designs rely on the dated Census frame for part of their sample. In practice, to obtain reliable survey estimates for most of the provinces, the LFS sample for all three designs would rely on rotation groups of various vintages. However, we consider here the simpler hypothetical situation in which the LFS rotation groups to be used for YITS are current – that is, the current household composition for the dwellings is known.[6]

[6] This scenario is pertinent to the results in Section 5. For the more complex case of rotation groups of various vintages, mobility of the target population between the LFS and YITS interviews reduces the effective sample size and biases the YITS estimates of the leaver rate.

Given the apparent dependency between mobility and leaving high school, in estimating the sample size for each of the dual-frame designs we should explicitly take account of the effect of mobility on the survey estimates. That is, we have to consider the distribution of leavers and movers in the target population at the time of the YITS collection.

We define the retrospective mobility function $f(\cdot)$, where $f(t)$ is the proportion of the population that has moved within the most recent time period t, that is, the proportion that has a different place of residence now than t time units ago. The function $f(\cdot)$ would be expected to vary according to age, sex and other characteristics of the population. In the case of YITS, we are interested in the mobility of the leaver domain, say $f_l(\cdot)$, within each of the province-sex target populations. In particular, we need $f_l(T)$, where T is the number of months between the 1996 Census and YITS data collection.

Let P be the leaver rate in the target population N to be surveyed by YITS. At time T the proportions of the target population comprised of leaver movers and leaver non-movers are respectively

$$P_M = P\cdot f_l(T) \qquad (4.1)$$
$$P_{\bar{M}} = P\cdot\left(1-f_l(T)\right)$$

with domain totals $Y_M = NP_M$ and $Y_{\bar{M}} = NP_{\bar{M}}$. As illustrated in Section 3, for the dual-frame design the domain of leaver movers is estimated only through the LFS sample, whether the screening estimator or the full-dual estimator is used. The domain of leaver non-movers is covered either by the Census sample alone or by both samples, depending on whether the screening estimator or the full-dual estimator is used. For the single-frame design based on only LFS groups, the leaver domain is estimated without distinguishing between movers and non-movers. The leaver non-mover domain can be estimated without bias by the Census sample. Similarly, the assumption that all LFS groups in the YITS sample are current implies the leaver domains estimated from this part of the sample are unbiased. Thus the estimators $\hat{Y}_{SCREEN}(d)$ and $\hat{Y}_{MULT}(d)$ are both unbiased for Y, the number of high-school leavers in the target population at Cycle 1 of YITS.

4.2 Feasible sample sizes

The target level of data quality is given by

$$CV(\hat{Y})=\alpha \qquad (4.2)$$

where $\hat{Y}$ is a dual-frame estimator of Y, that is, either the screening estimator $\hat{Y}_{SCREEN}(d)$ or the full-sample

estimator $\hat{Y}_{MULT}(d)$ defined in Section 3, where the domain d refers to leavers. In this context we can simply refer to $\hat{Y}_{SCREEN}$ and $\hat{Y}_{MULT}$. The sample size determination requires that we know the variance components of both the Census and the LFS frame. However, since we do not know these in practice, we approximate them by multiplying the simple random sampling variance by the design effect. Specifically we estimate

$$V(\hat{X}) \approx \frac{\beta(N-nr)N^2P(1-P)}{(N-1)nr} \qquad (4.3)$$

where N is the population size, X and P are the population total and mean for the characteristic of interest, β is the design effect for the sample design under consideration and n is the required initial sample size if we expect a response rate r. This assumes that non-response is random.

To estimate sample sizes for YITS, we rely on estimated leaver rates from the 1991 SLS, mobility rates for leavers and non-leavers estimated from the 1996 Census, and design effects for similar variables measured in other surveys based on the LFS and Census frames. We also assume the target populations covered by the Census and LFS frames are identical and of size N.

We first consider the estimator $\hat{Y}_{MULT}$. We want to determine sample sizes n_A and n_B from the Census and LFS frames, respectively, which satisfy the data quality constraint (4.2). Therefore

$$(\alpha NP)^2 = V(\hat{Y}_{MULT}) \qquad (4.4)$$

Substituting (3.2) and (3.3) in (4.4) and solving for $V(\hat{Y}_{A_{\backslash\bar{M}}})$, we obtain

$$V(\hat{Y}_{A_{\backslash\bar{M}}}) = \frac{Cov(\hat{Y}_{B_{\backslash\bar{M}}}, \hat{Y}_{B_{\backslash M}})^2 + V(\hat{Y}_{B_{\backslash\bar{M}}})\cdot\delta}{V(\hat{Y}_{B_{\backslash\bar{M}}}) + 2Cov(\hat{Y}_{B_{\backslash\bar{M}}}, \hat{Y}_{B_{\backslash M}}) - \delta} \qquad (4.5)$$

where $\delta = (\alpha NP)^2 - V(\hat{Y}_{B_{\backslash M}})$

Applying the variance approximation in (4.3) and the analogous equation for the covariance term, (4.5) may be rearranged to express the Census sample size n_A as a function of the LFS sample size n_B and various population and design parameters. The sample sizes depend on the population parameters N, P_M and $P_{\bar{M}}$; the design effects β_A and β_B for the leaver characteristic associated with the samples from the Census and LFS respectively; and the expected response rates r_A and r_B. Thus, for the full-dual estimator $\hat{Y}_{MULT}$, we can calculate pairs of feasible sample sizes (n_A, n_B) that would provide estimates of the specified reliability.

For the screening estimator $\hat{Y}_{SCREEN}$, feasible sample sizes (n_A, n_B) are also determined using the constraint (4.2), but in this case (4.5) is replaced by the much simpler relationship

$$V(\hat{Y}_{A_{\backslash\bar{M}}}) = (\alpha NP)^2 - V(\hat{Y}_{B_{\backslash M}}) \qquad (4.6)$$

Finally, for the single-frame design based on multiple rotation groups from the LFS, feasible sample sizes n_B are estimated by applying the approximation in (4.3) to the constraint (4.2).

4.3 Discussion of sample size estimates

To implement the dual-frame design for YITS, we would consider using only the LFS groups that were either current or had rotated out more recently than May 1996, the date of the 1996 Census. With the first cycle of YITS planned for January 2000, a maximum of 48 LFS groups (six from the December LFS and the rotate-out group from each of the preceding 42 months) would be available if none were taken as a sample by other longitudinal surveys.

As previously noted, the sample sizes presented here correspond to the simplified scenario in which all LFS groups used in the YITS sample are up-to-date. However, the limit of 48 rotation groups was still retained as a benchmark in computing the sample size estimates.

The population leaver rates and mobility rates applied in the sample size calculations were based on data from the 1991 SLS and the 1996 Census. The other important parameters include design effects and response rates. Design effects have been estimated for variables collected by other surveys that use the LFS frame. Also, the minimal degree of clustering in the Census sample suggested the design effect should be close to 1, and would be smaller than the design effect for the LFS sample. The values of 1.25 and 2.0, for the Census and LFS respectively, were applied in the sample size calculations. As for response rates, recent information from the LFS and other telephone surveys suggested there might be a response burden effect for households that had been in the LFS. The values of 0.80 and 0.70 were chosen for the Census and LFS samples respectively, although we hoped these would be conservative in practice.

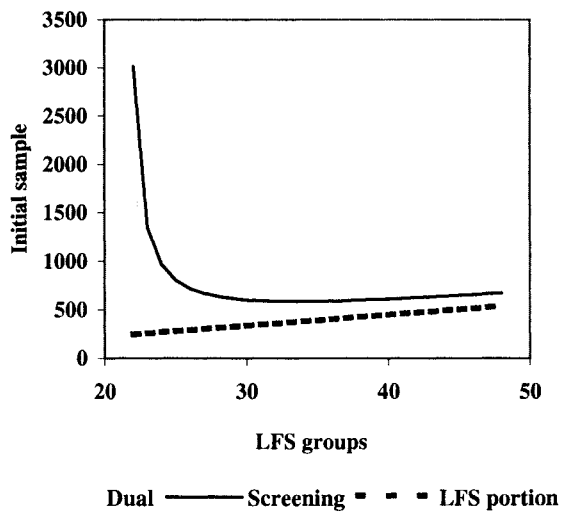
Ideally the target level for the CV of the estimated number of leavers in the target population would be about 16.5%. This is a benchmark Statistics Canada sometimes uses in advising data users on the quality of survey data. However, the scarcity of dwellings with members in the YITS target population, the relatively

small proportion of leavers and the high mobility of the population (and leavers in particular) make this level of data quality unattainable for many of the province-sex populations. This is particularly true for the Atlantic provinces, which have small populations, and for the female populations in many provinces, which have lower leaver rates than the males. Thus, sample sizes were computed based on CV targets of 20% and 25% for males and females, respectively.

Cost estimates for the total interview time were also computed for the feasible sample sizes associated with each design. We assume the front end of the interview takes an average of 10 minutes to administer, with 45 minutes required for the content modules. Thus for the dual-frame design based on the screening estimator, only the cost of the front end of the interview would apply to non-mover respondents in the Census sample.

In general, as a function of the number of LFS groups used in the YITS sample, the total sample size and total interviewing cost for the full dual and the screening estimator behave as indicated in Figures 4.1 and 4.2.

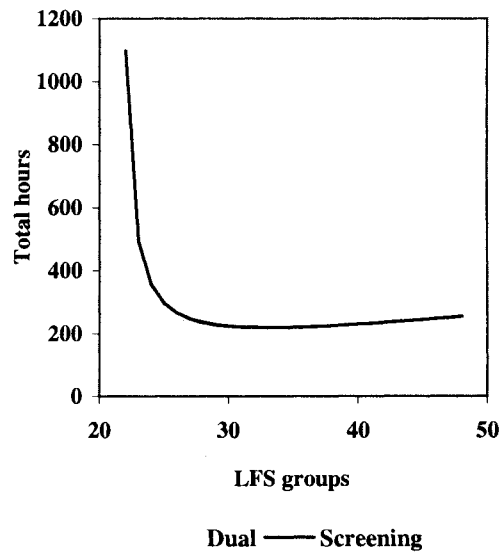
Figure 4.1 Sample size, male 20-year-olds, B.C.



By the nature of the full-dual design, feasible pairs of sample sizes are found on a continuum of decreasing values of the frame share λ. The largest values of λ correspond to a very small LFS sample and a large sample from the Census. As λ decreases the Census sample decreases and the LFS sample increases; the value $\lambda = 0$ corresponds to the single-frame LFS design. For a given number of LFS groups, the screening estimator requires a larger total sample size than that of the full-dual estimator, and the ratio of these sample sizes increases as the number of LFS groups increases, at least up to the number of groups required for the LFS single-frame. This seems logical because the addition of one more rotation group provides both mover and non-mover interviews for the full-dual design, but only mover interviews for the screening design.

Not surprisingly, for Ontario and Quebec, the two largest provinces, the single-frame LFS design might be acceptable given that no more than 12 rotation groups are needed. However, for the remaining eight provinces, the two dual-frame designs provide options to reduce the number of LFS groups used in the YITS sample, say by 20% to 30%, as long as one is prepared to increase the total initial sample size.

Figure 4.2 Interview hours, male 20-year-olds, B.C.



The fact that non-movers in the LFS sample are given a complete interview in the full-dual design and are asked only the front end of the questionnaire in the screening design has a large influence on the total costs of the two designs. The cost ratio of the screening to the full-dual design decreases as a function of the number of LFS groups. Among the feasible sample sizes for the dual-frame designs, the total interviewing costs almost always exceed those of the single-frame design. However, for most populations, the absolute cheapest option is the screening design with an LFS sample somewhat larger than that required for the single-frame.

5. Pilot survey

5.1 Design

In parallel with the estimation of sample sizes required for the dual-frame design, a national pilot survey was conducted early in 1999.

From the perspective of the dual-frame sample design, the most important objective of the pilot was to assess the response accuracy to the question on mover status. Given the degree of transition in various aspects of their lives, there was some doubt that respondents in the 18-20 age group would relate to the concept of usual place of residence or would accurately recall where they had been living 31 months earlier, at the time of the Census.

The sample design implemented for the pilot was, for the most part, the dual-frame design described in Section 3. It comprised a sample of 4024 dwellings, 2511 from the 1996 Census and 1523 from two LFS rotation groups, one group currently participating in the LFS, the other a rotate-out from May 1998. The size of the sample was determined in part by the intention to conduct the pilot longitudinally. To maximise the use of the sample for the pilot survey, two departures from the dual-frame design were implemented. First, a household member in the target population was eligible to be interviewed regardless of the stated mover status and the frame from which the dwelling had been selected. Secondly, if a person in the target population had been living in the selected dwelling and had moved out between the time of the Census and YITS, an attempt was made to contact this person if the dwelling was otherwise ineligible.

The sample from the Census consisted of two types of dwellings. The majority (2011) were dwellings which, at the time of the Census, had contained one or more household members born in 1977 to 1979.[7] This sample was stratified by province and household characteristics.

The second part of the Census sample was extraneous to the dual-frame design. It consisted of 500 dwellings which had no household members born in 1977 to 1979, but had contained a household member born in 1971 to 1976 – these individuals turned 20 to 25 years of age in 1996. This small sample was included to see if there might be particular kinds of dwellings that tend to attract young adults as residents. Positive evidence of this phenomenon actually would have lent support to the use of the Census as a single frame.

5.2 Pilot survey results

For the part of the pilot sample from the Census frame a dwelling identifier permitted a direct match to the 1996 Census database. Data for the date of birth and sex of YITS respondents were matched to those variables for persons that were household members of the same dwelling on Census day. A large sub-sample of these was manually verified to compare the names of the YITS respondents with those on the Census questionnaire for the same dwelling.

The results showed that the information on mover status provided by respondents is frequently in error, especially for true movers and especially if the source of information is a household member other than the designated respondent. People who are really movers tend to be classified as non-movers. Among respondents in the Census sample who were true movers and who answered the questions for the front end of the YITS interview, 21% (of 43) responded as non-movers. The information offered by a contact person other than the respondent was even more prone to error – among a total of 51 true movers, 71% were incorrectly classified.[8] On the other hand, non-mover misclassification was a negligible 3%.

The dual-frame estimators defined in Section 3 were used to evaluate the effect mover misclassification would have on the estimated leaver rates from a dual-frame sample. The results illustrated here are based on a selection of the dual-frame feasible sample sizes computed in Section 4, and the same underlying mover-leaver distributions. We also retain the assumption that the entire sample from the LFS is from a current frame, rather than from a series of groups of various vintages. This assumption eliminates the bias in the leaver rate estimate that would occur due to the vintage of the LFS sample. The effect we measure is then attributable only to the misclassification of mover status.

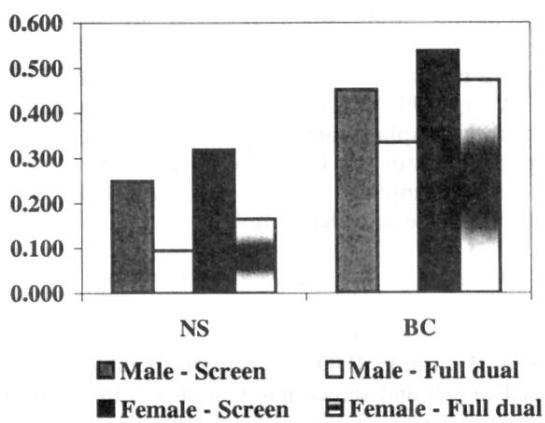
To compute the bias in the estimated leaver rate, we apply the probabilities of misclassification to each of the mover and non-mover domain counts expected in the Census and LFS samples. Within each of these domains, the probabilities of being misclassified and being a leaver are assumed to be independent. Due to the small sample size of the mover status results from the pilot survey, the probability of mover misclassification could not be evaluated for 20-year-olds by sex and province. Therefore the figures noted above for the entire Census sample in the pilot were applied to every province-sex population. Figure 5.1 illustrates the results for the two provinces with the smallest and largest relative bias, assuming a mover misclassification of 70%.

[7] The pilot survey date changed after the sample had been selected. Selected dwellings had household members born in 1977 to 1979, who turned 19 to 21 during the reference year 1998.

[8] Although the number of true movers in the Census sample was small, as expected, there was a contribution of 25 respondents from dwellings in the second part of the Census sample, all of whom were necessarily movers.

Under the stated assumptions, the misclassification of movers as non-movers causes a negative bias in the estimated leaver rate. For a given target population, the screening estimator is always more biased than the full-dual estimator. This arises because leaver movers in the LFS sample who are misclassified as non-movers are not represented at all in the screening estimator. In the full-dual design, however, these individuals are interviewed as non-movers and are represented by the share $(1-\lambda)$ in the estimated leaver rate. The magnitude of the relative bias varies considerably from one target population to another, but for the misclassification probability of 70% (associated with proxy information from the household contact), it is well over 10% for most of the populations, even for the full-dual estimator. In general the leaver rate is more biased for populations in which leavers are much more mobile than non-leavers – thus the relative bias for female populations tends to be larger than that for males. It is interesting to note that a misclassification of non-movers does not cause any bias in the estimated leaver rate. This is due to equal biases of opposite sign that occur in the non-mover components and the mover component of the dual-frame estimators.

Figure 5.1 Relative absolute bias of estimated leaver rate due to mover misclassification



6. Adopted sample design

With the pilot survey evidence on mover misclassification and the sample size comparisons among the alternative designs for YITS, the dual-frame design fell out of favour. The design implemented for the main survey relies solely on the LFS sample, that is, a combination of active and rotate-out groups.

In this single-frame design, persons in the target population (i.e. persons born in the years 1979 to 1981) were identified directly from the LFS roster and were traced, if necessary, starting with names, addresses and telephone numbers from the LFS. The selected individuals were to be contacted and interviewed regardless of their usual place of residence at the time of the YITS collection. Due to the expected effort of tracing individuals who would have changed dwellings since the LFS interview, the sample for YITS was limited to the current LFS sample for December 1999 and groups that had rotated out of the LFS since January 1997. A total of 36 rotation groups were available for YITS. Among the designated rotation groups a sample of approximately 29,000 persons was selected.

7. Conclusion

Cycle 1 data were collected from January to April 2000. As expected the tracing activities consumed a lot of resources but the effort paid off. The overall response rate was 81%; this excludes 3% of cases in the initial sample that, on contact, were found to be out-of-scope. At the end of collection only 7% of the entire sample was flagged as untraceable.

And what of the risk of non-response bias arising from sampling older LFS rotate-out groups? The data show a non-response rate increasing with the age of the rotation group, ranging from 15% for the six most recent groups to 22% for the six oldest groups. So the potential for non-response bias in the survey estimates does exist. A preliminary examination of the estimated leaver rate by vintage of the rotation group did not reveal an obvious bias. However, more work has to be done in this area to determine the nature of any non-response bias and approriate adjustments to be incorporated in the overall weighting strategy.

8. Acknowledgments

The authors wish to thank Geoffrey Hole and Johanne Denis for their helpful comments on an earlier version of this paper.

9. References

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. and Lindeyer, J. (1998). Methodology of the Canadian Labour Force Survey. Statistics Canada, Catalogue no. 71-526.

Skinner, C.J., and Rao, J.N.K. (1997). Estimation in Dual Frame Surveys with Complex Designs. *Journal of the American Statistical Association*, Vol. 91, p. 349-356.