THE NATIONAL POPULATION HEALTH SURVEY: DESIGN AND ISSUES

M.P. Singh, J.L. Tambay, S. Krawchuk, Statistics Canada
J.L. Tambay, 16-L R.H.Coats Bldg., Statistics Canada, Ottawa, Ontario K1A 0T6

KEY WORDS: Longitudinal Surveys, Rejective Method, Two-Phase Sampling

I. INTRODUCTION

Statistics Canada is conducting the National Population Health Survey (NPHS), a comprehensive longitudinal survey covering a variety of aspects related to health. The target population includes household residents in all provinces and territories, with the principal exclusion of populations on Indian Reserves, Canadian Forces Bases and some remote areas. A separate institutional component will cover long term residents of hospitals and residential care facilities.

The initial wave sample, in 1994, will provide a panel of respondents who will be followed-up every two years for up to 20 years. In this wave, data are being collected for the following general areas: a) Measures of health status including a health index, identification of health problems, measures of mental health and of functional limitations; b) Determinants of health which include behavioural, lifestyle, economic and social factors; c) Use of health services and prevention practices; and d) Socio-demographic information including educational attainment and labour force characteristics of household members. Most of the information in categories a) to c) is collected for only one person per household: the randomly-selected longitudinal panel member.

The production of provincial cross-sectional estimates is another main objective of the 1994 wave. The sample size, originally set at 22,000 households, was increased through provincial buy-ins to 26,000 to allow for sub-provincial estimates in some cases. The household sample will be split among four collection periods: June, August, November, and March 1995.

This document has seven sections. Section 2 presents the general design of the sample as it is used in most provinces. The section covers the allocation of the sample to the provinces and territories, the use of a rejective approach to improve sample representativity, and the sample selection. Section 3 covers the adjustments to the sample design following the decision to integrate the NPHS with the National Longitudinal Survey of Children, a survey being carried out in December, 1994 and February, 1995. Section 4 presents the sample design used in Quebec, where the NPHS design benefitted from household information from a provincial health survey. The

design in the territories is briefly presented in section 5. Section 6 gives the anticipated distribution of the sample by age and sex and section 7 discusses issues related to the design of the survey for future waves.

II. REGULAR SAMPLE DESIGN FOR HOUSEHOLDS

The sample allocation, the application of the rejective method, and the selection of the sample in provinces outside Quebec are covered in this section.

2.1 Sample Allocation

The NPHS was budgeted for a sample size of 22,000 households. It was further agreed among national and provincial/territorial representatives that each province and territory needed a minimum of 1,200 households. Subject to this restriction the provincial and territorial sample sizes were obtained by using a well known allocation scheme that balances the reliability requirements at national and regional levels (Kish, 1988). According to this scheme the sample was allocated proportionally to $\sqrt{(0.804W_h^2 + 1/12^2)}$, where W_h is the 1991 Census proportion of households in province/territory h, h=1,..,12. This allocation determined the base sample size for each province and territory. Greater sample sizes could be obtained through the buy-in of additional sample units.

Within provinces allocation was initially proportional to the population size. This approach was used for the following reasons: a) it was optimal for estimates of ratios and percentages; b) it allowed self-weighting designs which are simpler in weighting and to analyze; c) it presented a good compromise in designs where auxiliary information correlated with study variables was not available or where the multitude of characteristics studied were related to different sets of auxiliary variables; and, d) it could simplify the use of a multi-purpose design for the sample (see 2.3).

Four provinces decided to have additional sample to satisfy certain reliability criteria for specified sub-populations. These buy-in samples will not normally become part of the longitudinal sample.

Ontario bought additional sample in Health Regions to allow for estimations of given accuracy for 2 or 3 age/sex groups by Region. Manitoba increased the sample sizes to a total of 450 households in Winnipeg and 225 in other Health Regions. In both

provinces the sparsely-populated northern Regions were merged to keep buy-in costs down. New Brunswick bought sample to increase the allocation in less populated regions.

British Columbia requested a buy-in of 850 households strictly for the Health Unit covering Prince George. As the increase was too high to be accommodated by locally available interviewers most of the buy-in sample is to be covered by Random Digit Dialling (RDD). This was acceptable as there were no longitudinal requirements of the buy-in sample (RDD respondents are reluctant to give tracing information). The non-RDD portion was incorporated into the regular sample requirements.

In applying the rejective method, sample sizes everywhere were further inflated by the number of households expected to be screened out of the sample (see 2.2).

In the two territories it was decided to fully integrate the survey with the National Longitudinal Survey of Children (NLSC) to reduce response burden. A sample size of 1,500 households per territory was necessary to yield required samples of children for the NLSC. The NPHS would subsample its 1,200 households from among the 1,500 using a rejective method similar to that used in the provinces.

Table 1 below gives the sample sizes expected by province and territory. Numbers represent in-scope private occupied dwellings before nonresponse, which is anticipated to be near 10%.

TABLE 1: SAMPLE SIZES FOR THE NPHS

			Household sample sizes		
Prov./ Terr.	Orig. alloc.	Buy-in sample	To interv'w	Screened out	Total
Nfld	1,220	–	1,221	171	1,392
P.E.I.	1,201	–	1,199	223	1,422
N.S.	1,270	–	1,270	246	1,516
N.B.	1,243	180	1,423	234	1,657
Que.	3,584	–	3,479	–	3,479
Ont.	4,817	2,183	7,001	1,021	8,022
Man.	1,307	493	1,800	324	2,124
Sask	1,287	–	1,288	257	1,545
Alta	1,674	–	1,674	305	1,979
BC (1)	1,996	61	2,057	448	2,505
Sub-tot	19,599	2,917	22,413	3,229	25,642
BC (2)	–	788	788	–	788
Yukon	1,200	–	1,200	300	1,500
N.W.T.	1,200	–	1,200	300	1,500
TOTAL	21,999	3,705	25,601	3,829	29,430

(1) Excludes RDD portion. (2) RDD portion.

2.2 The Rejective Approach

The survey content primarily focuses on one member in each sample household who is chosen at random to become the longitudinal panel respondent. As the larger the household, the less chance that a particular member is chosen, the panel underrepresents persons coming from large households, typically parents and children, and overrepresents persons coming from small households, often single people or the elderly.

A rejective approach has been adopted to increase the representation of parents and youths in the panel. A portion of the sample is pre-identified for screening. After their member roster is completed, screened households that have no member aged under 25 years drop out of the survey. In order to maintain the required sample sizes, the number of households visited in each province is increased by the anticipated number of households screened out in this way.

As applying a rejective method increased the collection costs and introduced some variability in sample sizes it was only adopted after cost-free alternatives had been considered. One possibility was to increase the chance of selecting for the panel certain person types, for example the household children. However, as under-representation was an issue for large households, this only shifted the problem to other members of these households like the parents. Another possibility was to increase the sample representation in areas where there were higher concentrations of large households. However, aside from apartment buildings, households were not sufficiently concentrated by size to yield satisfactory results.

The rejective method with an under 25-year old rule was adopted as it performed better than other rejection rules considered (under 20-year old rule, 3-members minimum rule, and a combination of both). For cost and operational reasons the percentages of screened households was usually limited to 25-30% in Ontario, 37.5-40% in urban areas elsewhere and 25-30% in rural areas. As apartment strata had a high concentration of small households, their sample sizes were reduced instead of applying a rejective method. The rejective approach was not applied in remote regions either because of the high contact costs there, and its use was limited in areas where sample buy-in demands were substantial.

2.3 Sample Selection

The sample design considered for the household component of the NPHS was a stratified two-stage design. In the first stage homogeneous strata are formed and independent samples of clusters are drawn from each stratum. In the second stage dwelling lists are prepared for each cluster and dwellings, or households, are selected from the lists.

In all provinces except Quebec the NPHS used the multi-purpose sampling methodology developed for the redesign of the Labour Force Survey (LFS). That methodology provides general household surveys with clustered samples of dwellings, thus making the design very cost effective for the listing and collection of data. The next three paragraphs describe the design of this multi-purpose sample.

The basic design is a multi-stage stratified sample of dwellings selected within clusters. For design considerations, each province is divided into three types of areas: Major Urban Centres, Urban Towns and Rural Areas. Geographic and/or socio-economic strata are formed within the Major Urban Centres. Some Centres have separate Apartment frame strata and strata of Census Enumeration Areas (EAs) with high average household incomes. Six clusters or apartment buildings (sometimes 12 or 18) are selected from each stratum using a randomized probability proportional to size (PPS) sampling scheme, size being their number of households. The number six is used throughout the sample design to allow a one-sixth rotation of the sample every month for the LFS.

Remaining urban towns and rural areas in each province are stratified first by geography then by socio-economic characteristics. In most strata six clusters, usually Census EAs, are selected with PPS. In a few cases where the population density is low, a three stage design is obtained by first selecting 2 or 3 PSUs (Primary Sampling Units, usually groups of EAs) and then dividing each PSU into clusters, six of which get sampled. Selection at each stage is done with PPS.

The sample of dwellings is obtained after listing operations in sample clusters are completed. As sampling rates are predetermined there are often differences between anticipated and obtained sample counts. Excessive sample yields are corrected by dropping a portion of the originally selected units. This is usually done at aggregated levels and is called sample stabilisation. Note also that sample sizes are inflated to represent dwellings rather than households as approximately 15% of the dwellings are expected to be vacant or otherwise out-of-scope.

The sample design is set up to yield about 60,000 households. Surveys needing smaller sample sizes usually "reserve" from 1 to 6 rotations per province, a rotation being one-sixth of the total sample. Sample stabilisation is used to maintain the sample at desired levels, as when two rotations are reserved but the sample size needed only represents 1.5 rotations.

This sampling strategy could not meet NPHS sample needs without modification. Firstly, as a result of sub-provincial buy-ins and other factors the design did not reflect NPHS sub-provincial allocation needs. Asking for a fixed number of rotations throughout a province was inadequate in some regions and inefficient in others. Secondly, NPHS needed to satisfy additional sample requirements at the cluster level. For variance estimation sample clusters in each stratum had to be divided into 2 or more replicates (i.e., subsamples which are selected in an independent and identical manner). The sample had also to be distributed among the 4 collection periods but, to reduce costs, it was better to visit each cluster in one collection period only. The number of clusters selected per stratum thus needed to be 8 or a higher multiple of 4.

The first requirement was met by allowing different numbers of rotations to be requested within a province. The second requirement was equivalent, in most cases, to having a number of rotations which was a multiple of 4 (rotations were independently identically distributed replicates except in strata with a three stage design, where replicates were the PSUs). As NPHS usually requested only between 2 and 6 rotations it was necessary to form groups of "similar" strata to meet this requirement. Factors considered when forming groups were the sector (Major Urban, Urban Towns or Rural), geographic proximity and, in large cities, average household income. When strata were grouped, their sample clusters were also grouped to form replicates.

As a result of these modifications, the NPHS sample of clusters can be considered as a stratified replicated sample where strata are groups of the original strata and replicates are typically independent, identically distributed samples of 4 clusters each. There were exceptions, but they are not expected to have a significant impact on survey results.

III. INTEGRATION WITH THE NATIONAL LONGITUDINAL SURVEY OF CHILDREN

The National Longitudinal Survey of Children (NLSC) is a household survey which will follow a sample of about 25,000 children under 12 over time. It is being integrated with the NPHS as the surveys are interested in similar characteristics for children. In the territories the surveys will use a common household sample and questionnaire. This section deals with integration in the provinces, where it is limited to the collection of common data for children and the use of a common Computer Assisted Personal Interview application. In order to provide a sample of 4,500 to 5,000 children to the NLSC there, NPHS households where a child is selected for the panel will have detailed children's questionnaires administered to all children present (subject to a maximum of 4).

Scheduling constraints require that children not be selected for the NPHS panel before the third collection period (or quarter). This will distort the seasonal representativity of children in the panel and reduce their sample size. To increase the sample yield for children without harming the seasonal representation of other household members in the last two quarters it was decided to reassign part of the NPHS sample from the first two quarters to these quarters. As this decision was made after the sample operations described above, the shift is applied to households within clusters rather than to entire clusters.
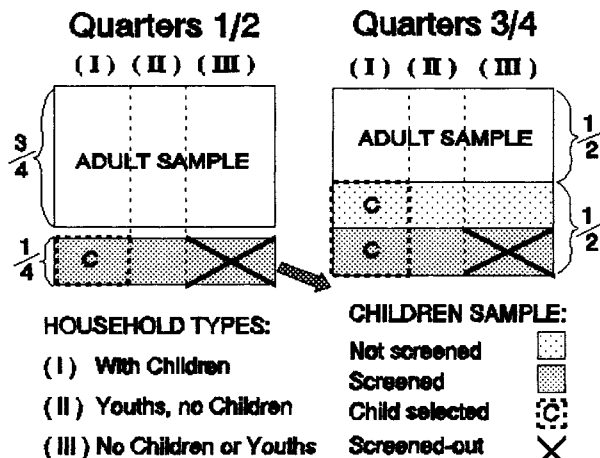
Figure 1: Repartition of the Sample

Figure 1 illustrates how the sample distribution was revised for the integration. The square on the left represents a cluster assigned to quarter 1 or 2. That on the right represents a cluster assigned to quarter 3 or 4. Households are classified by type into: (I) households with children; (II) other households with youths ("Youth" meaning under 25 years old); and (III) households without children or youths. The sample is divided into an "Adult" sample and a "Children" sample. In "Adult" sample households only persons aged 12 or older can be selected for the panel. Panel membership is restricted to children in "Children" households. If there are no children present, then either the household is screened out or a member (non-child) is selected at random for the panel.

A quarter of the sample from quarters 1 and 2, and a half from quarters 3 and 4 are designated as "Children" households. "Children" households from quarter 1 or 2 will actually be visited in quarter 3 or 4, respectively. Outside of Prince Edward Island the rejective method can be applied strictly within the "Children" sample. When the screening rate is at 37.5% all "Children" households are screened. With lower rates some of them will not need to be screened. A 25% screening rate is illustrated in Figure 1. All the "Children" households from quarters 1 and 2 and one-half of those from quarters 3 and 4 are screened.

With this method the number of non-children in the panel will be approximately the same in each quarter. However, there will be seasonal differences in sample yields within each household type. In Type I households (those with children) 50% more non-children will be interviewed during the first two quarters because "Adult" households constitute 3/4 of the sample in quarters 1 and 2 and only 1/2 in quarters 3 and 4. The shifting of the "Children" sample to the last two quarters also means that in Type II households (youths but no children) there will be 67% more non-children selected in the last two quarters. For Type III households (no children or youths) the seasonal distribution will vary according to the screening rate. With a 37.5% screening rate results will be the same as for Type I households, while with a 25% screening rate the number of non-children selected will be the same throughout the year.

For operational reasons there are no rejections and no shifting of collection periods in apartment strata, high income strata and remote regions. In Prince Edward Island the number of available interviewers did not permit shifting the collection periods, and screening occurs in all quarters. The "Children" sample in these cases is selected strictly from clusters in quarters 3 and 4, resulting in a seasonal distortion of the sample for non-children.

IV. SAMPLE DESIGN IN QUEBEC

In Quebec the NPHS sample is selected from dwellings participating in a health survey organized by Santé Québec: the 1992-93 *Enquête sociale et de santé* (ESS). This was mutually beneficial as Santé Québec would obtain longitudinal coverage for households agreeing to share their NPHS data, and NPHS could use ESS data to improve the representativity of their sample without having to screen out households.

The ESS covered 16,010 dwellings selected using a two-stage design similar to that of the LFS. The province was divided geographically by crossing 15 Health Regions with four urban density classes (Montreal Census Metropolitan Area, regional capitals, small urban agglomerations and the rural sector). In each area clusters were stratified by socio-economic characteristics and selected using a PPS sample. Selected clusters were enumerated and random samples of their dwellings were drawn: 10 per cluster in major cities, 20 or 30 elsewhere.

Santé Québec provided non-confidential information which allowed the classification of their sample into 4 types of households: one-member households; households with children; other households with youths (persons aged under 25); and the rest (more than one member and no youth or child). ESS nonrespondents were imputed a household type by the NPHS.

The NPHS sample size was first allocated among the four urban density classes. To avoid having too much sample in Montreal the allocation was proportional to $\sqrt{(2W_h^2 + 1/4^2)}$, where W_h is the population share for class h, h=1,2,3,4. In each class an attempt was made to obtain a subsample from the ESS which, as far as the selected panel member was concerned, would be proportional to the populations for the 4 household types. This was done by drawing a sufficient number of households from the ESS to give the required yield for households with children (the most underrepresented group), and then removing excess sample from the other three household groups. An initial sample which was almost 50% higher than needed was thus selected. After removing from it 2/3 of the one-member households, 1/2 of the other households with no youths or children, and 1/6 of households with youths but no children, the objective was nearly attained.

Considerations for seasonal representation and variance estimation, and integration with the NLSC, affected the sub-sampling in Quebec as they did elsewhere. ESS strata were thus collapsed to allow the formation of replicates, with the clusters in each replicate covering all four quarters (two quarters are covered per cluster in the rural and small urban sectors as sample sizes are higher there). The sample of households with children was split into an "Adult" sample and a "Children" sample by a 3:2 ratio, the terms having the same meaning as in other provinces. "Children" sample households in quarters 1 and 2 were reassigned to quarters 3 and 4. As NPHS surveys the current occupants of dwellings selected for the ESS, and changes will have occurred in some of those dwellings, the samples of households without children for quarters 3 and 4 are also to be split, by a 2:3 ratio, into an "Adult" and a "Children" sample.

The table below gives the expected distribution of the sample based on ESS data. The slight overallocation in one-member households was intentional. This group has higher nonresponse rates and is the most likely to increase in size as household compositions change.

TABLE 2: QUEBEC SAMPLE DISTRIBUTION BY HOUSEHOLD TYPE

HOUSEHOLD TYPE	% POP.	SAMPLE	
One-member households	9.2%	415	(12%)
Households with children	37%	1,293	(37%)
- Child selected	15%	526	(15%)
- Non-child selected	22%	767	(22%)
Others with Youths	29%	1,033	(30%)
Others, no Youths or Children	24%	738	(21%)

V. SAMPLE DESIGN IN THE TERRITORIES

In the territories the NPHS and NLSC will be conducted as a single survey in November 1994 to reduce the respondent burden. A sample of 1,500 households, including 300 to be screened out, will be selected in each territory to allow a sufficient yield of children for the NLSC. This assumes that children present in sample households will be covered for the NLSC (subject to a maximum of 3 per household).

In the Northwest Territories households will be selected randomly from each non-remote community above a certain size. In the Yukon households will come from a PPS sample of larger communities.

VI. ANTICIPATED DISTRIBUTION OF THE SAMPLE BY AGE AND SEX

Table 3 below gives expected distributions by age and sex for the total of the provincial samples excluding the RDD buy-in sample in B.C. Figures are approximate because the sample design, the rejective method, the respondent selection, and finally survey nonresponse all introduce variation in the results. Nonresponse may cause a drop of about 10% in the sample yields given. Two sets of figures are given. Full sample results are based on the full composition of sample households and panel sample results are based on the distribution of the selected respondents to the panel (one per household). The full sample figures are given because some of the survey questions are administered to all household members.

TABLE 3: ESTIMATED SAMPLE YIELD BY AGE AND SEX

	Under 12	12-24	25-44	45-64	65 & ovr	TOTAL
A: Full Sample Results (all household members)						
Male	6,513	6,988	10,657	5,851	2,491	32,500
Female	6,199	6,811	11,294	5,655	3,119	33,078
TOTAL	12,712	13,799	21,951	11,506	5,610	65,578
B: Panel Sample Results (1 member/household)						
Male	1,458	1,982	3,824	2,246	1,298	10,808
Female	1,381	2,022	3,914	2,339	1,967	11,623
TOTAL	2,839	4,004	7,738	4,585	3,265	22,431

The age distribution of the selected respondents has been helped by the rejective method outside Quebec and by the special design in Quebec. The expected number of children interviewed for the NLSC will be 4,746 (2,434 boys and 2,312 girls). It is higher than the number for the panel because the NLSC will cover all the children in "Children" households up to a maximum of 4 in the provinces.

VII. DESIGN ISSUES FOR FUTURE WAVES

In considering future waves for this biennial survey the sometimes conflicting objectives of longitudinal and cross-sectional estimation are raising more design issues. The longitudinal panel, although able to produce good cross-sectional estimates in 1994, would become inadequate for cross-sectional estimation after only a few years. By 2004 over 15% of the Canadian population will have been born or immigrated since 1994. These persons will not be covered by the 1994 longitudinal panel.

Approaches for topping-up the sample for cross-sectional estimation were explored. They were based on a fixed sample of 22,000 persons per wave, that is, assuming extra sample could only be selected up to the level that was lost through attrition. Assumptions were made for the numbers of births and deaths, immigrants, nonresponse, movements out-of-scope and unable-to-trace situations.

Under the above scenario the sizes of the top-up samples from 1996 to 2004 would be of 5,376, 3,210, 2,535, 2,366 and 2,327 persons, respectively, after nonresponse. Although these numbers were sufficient to cover the percentage of the population born or immigrated since 1994 (assuming that these persons could be sampled, say, from LFS rotated-out dwellings) there would still be the problem of overall representativity of the sample as it aged, moved or became nonrespondent.

An alternative considered consists of topping-up the sample every four years for cross-sectional purposes. The in-between waves would be used primarily for longitudinal estimation. Parts of the top-up sample would be kept in the panel while the rest would only be used once. As the top-up would not be used half of the time, the sample size could be increased beyond the 22,000 units in the years that a top-up was used. Results of the 1996 wave (a non top-up year) would be used to decide how best to identify the top-up sample in 1998.

Another possibility consists of following only a portion of the sample if they move in addition to keeping all of the dwellings originally selected for cross-sectional purposes. If one-third of the 25-30% of people that move in two years are traced then the sample size would increase by 8-10% in 1996, and 80-83% of the original panel would remain in the sample. A variant which maintains the sample size would consist of splitting the sample into a "household" and a "dwelling" portion. The panel members in the "household" portion would be traced and interviewed, with their household, in 1996 and beyond. Only current occupants of the "dwelling" sample would be interviewed each wave. If the sample is split 50-50 then 12-15% of the panel would be lost each wave.

The relative merits of "household" versus "dwelling" approaches are still under consideration. Apart from the obvious advantage for longitudinal estimation, a "household" approach is cheaper when future waves are followed by telephone and it allows coverage of newly constructed areas. A "dwelling" approach, used for cross-sectional estimation, requires no tracing and is simpler for estimation. It also has more stable sample sizes by region. When the population of movers has distinct characteristics then both methods hurt the panel representativity. A "dwelling" approach loses longitudinal coverage of the young and single adults while a "household" approach loses cross-sectional coverage of economically impoverished regions.

The final methodology adopted would benefit from a classification of movers into two types. One-half of the population that had moved between the 1986 and 1991 Censuses had stayed in the same municipality while another third had moved within their province. The first group, and perhaps the second, do not present regional coverage problems for cross-sectional estimation.

ACKNOWLEDGMENTS

REFERENCES

Catlin, G. and Will, P. (1992). The National Population Health Survey: Highlights of Initial Developments. *Health Reports*, Vol. 4, No. 3, Statistics Canada, Cat. 82-003, 313-319.

Kish, L. (1988). Multipurpose Sample Designs. *Survey Methodology*, 14, 19-32.

Singh, M.P., Drew, J.D., Gambino, J.G. and Mayda, F. (1990). *Methodology of the Canadian Labour Force Survey 1984-1990*, Statistics Canada, Cat. 71-526.

Statistics Canada. (1992). *Age, Sex and Marital Status - The Nation*, Cat. 93-310.

Health and Welfare Canada and Statistics Canada. (1993). *National Longitudinal Survey of Children (NLSC) Overview*.