IMPROVING SAMPLE REPRESENTATIVITY THROUGH THE USE OF A REJECTIVE METHOD

J.L. Tambay, C. Mohl, Statistics Canada
J.L. Tambay, 16-L R.H. Coats Building, Statistics Canada, Ottawa, Ontario K1A 0T6

KEY WORDS: Longitudinal Surveys, Respondent Selection

1. Introduction

The National Population Health Survey (NPHS) is a longitudinal household survey that will follow a panel of respondents every two years for up to twenty years. General health and socio-demographic information will be collected for all members of sample households while detailed information on health status, determinants of health, health prevention practices and other related subjects will be collected from one household member only: the randomly selected longitudinal panel respondent. Panel respondents are selected within sample households in the first wave. Households in future wave samples will be those in which the original panel respondents live.

By selecting only one member of the household to be a longitudinal member, people living in smaller households, for the most part single people and the elderly, have a higher probability of being selected than those in larger households, where most parents and children live. This is an undesirable property since the composition of the longitudinal panel should be representative of the population as a whole. To reduce the effect of this poor representativity, the NPHS uses a rejective method of sampling where a household may be rejected, that is, dropped from the sample, if the members do not have certain characteristics.

This report examines how the NPHS rejective method is used to improve the representativity problems in the longitudinal sample and how this method affects results. Section 2 explores the advantages and disadvantages of selecting one person per household compared to selecting all members. Section 3 looks at the NPHS design and the methods that were considered to make the sample more representative. The results of applying the rejective method to the NPHS are also shown. Section 4 examines the properties of the rejective method from a theoretical point of view. The effect of the method on variance and sample size is examined. An illustration of the impact of the rejective method is shown in Section 5. Some conclusions are presented in Section 6.

2. The One Person Per Household Rule

The decision to focus on one household member was made to allow more in-depth questioning of the selected member without placing excessive respondent burden on sample households. Some information for which it is better to use a larger sample, such as on the prevalence of health conditions or the utilisation of health services, is still collected from all of the people living with panel respondents at each wave. Focusing on one member also simplifies longitudinal follow-up operations, since procedures to deal with household changes that occur over time are not necessary.

Statistics Canada's National Health Promotions Survey[1] and General Social Survey[2] interviewed one person per household. Both were conducted using random digit dialling. New Zealand's Household Health Survey[3] also selected only one household member for in-depth interviewing.

Alternatively, many other health surveys, such as the 1978-79 Canada Health Survey[4], the 1990 Ontario Health Survey[5], Quebec's 1992-93 *Enquête sociale et de santé*[6] and the United States' National Health Interview Survey[7] have chosen to interview all household members. Aside from the obvious advantage of using a larger sample base, these surveys also avoid some of the sampling issues that come from subsampling within households, namely the loss of precision due to differential weighting and a non-representative sample distribution. Differential weighting occurs because the sample weights fluctuate according to the household size. The loss of representativity results when households are not selected with probability proportional to household size - which is often the case when household compositions are not known prior to collection. Since persons coming from small households have a greater chance of being retained in the panel than persons coming from large households, the panel under-represents the latter, which tend to be parents and children, and over-represents the former, typically single persons and the elderly.

The rejective method helps to correct problems of representativity which cannot generally be treated before collection. The following sections illustrate how and why the method was used for the NPHS.

Before concluding this section, some disadvantages of interviewing all household members are given for the sake of completeness. One issue already covered is the greater respondent burden imposed on sample households (or, alternatively, less information is collected in the interest of maintaining an acceptable level of response burden). If the survey contents do not allow for proxy interviewing, this can also translate into more visits to the household being necessary to complete the interviews. Also, the presence of intra-household correlation (members share common socio-demographic and economic characteristics) means that less information is collected from the sample than if respondents had come from a greater number of households. Finally, for longitudinal surveys, some of the undesirable properties of subsampling within households, such as the presence of differential weights, are unavoidable as the household compositions evolve.

3. The NPHS and Methods to Correct its Representativity Problem

In this section the general design of the NPHS is presented along with an illustration of the representativity problem that is encountered by using such a design. Advantages and disadvantages of methods which were considered to help correct the representativity problem are also examined.

3.1 Use of the LFS design for the NPHS

In nine of the ten provinces (excluding Quebec) the NPHS used the general household sample selection methodology developed for the redesigned Canadian Labour Force Survey (LFS). The LFS uses a deeply stratified multiple stage sample design which is suitable for many types of household surveys. NPHS households are selected from households about to be rotated into the LFS. Selection of these households by the NPHS precludes them from future coverage by the LFS. No prior information is known about their composition other than the basic LFS design information (strata characteristics). One member is chosen from each responding household to be the longitudinal respondent.

By selecting only one person from each household to receive the detailed health questionnaire, certain age groups are greatly under-represented while others are over-represented. Table 1 compares the age breakdowns of the in scope population from the 1991 Census to the breakdowns from the LFS sample where every member is selected and a simulated NPHS sample selection where only one member of an LFS household is chosen.

Table 1
Percent Distribution of Sample by Age Group

Age Group	1991 Census	LFS Sample	NPHS Sample
0-4	6.99	7.23	4.90
5-9	6.90	6.85	4.58
10-14	6.83	6.84	4.53
15-19	6.77	6.68	4.94
20-24	7.33	7.14	6.76
25-34	17.93	17.17	17.03
35-64	36.24	36.90	38.79
65+	11.00	11.20	18.48

Clearly the simulated NPHS sample severely under-represents children and youths while grossly over-representing seniors. Young adults are also under-represented.

3.2 Alternative Methods for Improving the Representativity of the NPHS Sample

A number of methods were examined in order to find a way to improve the representativity of the NPHS sample. In addition to improving the representativity, the chosen method also required the following properties.

1) It had to keep response burden to a suitable level.
2) It had to be easy to implement using the NPHS collection methodology (four collection periods per year).
3) It could not increase the problems of differential weighting.

3.2.1 Units Rotated Out of Other Surveys

If a household has responded to another survey such as the LFS, the dwelling composition is known. Households may then be selected in such a manner to increase the number of longitudinal members in under-represented groups. The problem is that the burden put on the respondents may result in higher non-response than one would get from a fresh sample. The Quebec sample of the NPHS used units previously surveyed in the *Enquête sociale et de santé* (ESS) as this presented advantages both to the NPHS and to Santé Québec, which organized the ESS. It was judged preferable to use a fresh sample elsewhere.

3.2.2 Changing the Individual Probability of Selection

Another option is to increase the probability of selecting people in the under-represented groups within sample households. This can be achieved by giving each member a 'weight' based upon which group he/she falls into. An individual's selection probability is then based upon the weight rather than

being equal for each member. While this may improve the representativity of the sample by age, there are still the same number of people coming from small and large households as before. It has already been noted that most larger households consist of children and their parents. Thus, increasing the probability of selection, say, for a child, results in a smaller chance of selecting a parent. The parents, already under-represented in the equal probability of selection sample will become even more under-represented. In addition, the problem of differential weighting also increases since the parents now have an even smaller probability of selection. This means that, if selected, they will have an even larger weight compared to an equal probability of selection scheme.

3.2.3. Changing the Allocation of the Sample

A third option is to increase the number of large households in the sample and hence select more children and parents as longitudinal respondents. This can be achieved by assigning a greater proportion of the sample to strata with a greater number of large households. A simulation was done using LFS data since the LFS is the basis for the NPHS sample. The results were not encouraging since, with the exception of the apartment strata, the LFS strata were not homogenous enough with respect to household size for this method to work effectively.

3.2.4 Two Phase Design

Using this method, a large sample of households is initially visited and the demographic characteristics of all of the members are collected. The results from the sample are then pooled together, and a subsample of households is selected. The subsample is chosen so that, after selecting a household member, it is representative of the population as a whole. This method requires extra costs since a larger sample is needed for the first phase and those households which are selected for the second phase have to be revisited. Since the NPHS methodology required sample to be selected at four points throughout the year, this method was also not feasible as it would have required all of the first phase information to be collected before the second phase could commence. In Quebec, the sample is effectively a two-phase sample where phase one consisted of dwellings selected in the *Enquête sociale et de santé* sample.

3.2.5 Rejective Method

This is the method that was finally adopted for use in the NPHS. It is similar to a method of sample selection used in the United States' National Health Interview Survey[7].

Figure 1
The Rejective Method

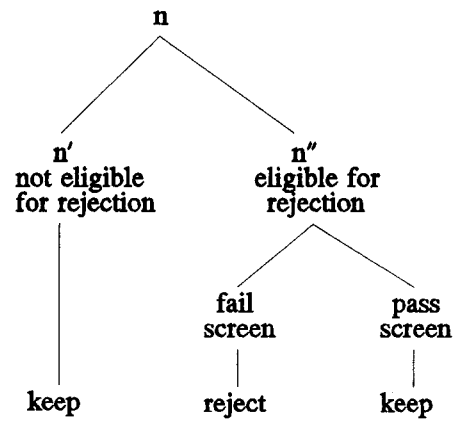


Figure 1 demonstrates how the rejective method works. An initial sample of size n is selected. A subsample of size n" is pre-identified as eligible for rejection (EFR). This pre-identification is done independently within each replicate. Upon visiting an EFR dwelling, the interviewer administers a screening questionnaire that determines the composition of the household. If the household does not have the characteristic of interest (in the case of the NPHS, if it does not contain anyone under the age of twenty-five), then the household is rejected. If the EFR household has the characteristic of interest, or if the household is not eligible for rejection, then the interview continues in a normal fashion.

As in the case of two-phase sampling, a larger initial sample size is required for the rejective method. However, the interviewer only has to visit the household once. Although the sample size is variable, since the number of units that will ultimately be rejected is unknown, it is more stable than using a Bernoulli trial to determine if a dwelling without the characteristic of interest should be dropped. Further investigation of the sample size stability can be found in Section 4.3.

For more information on the design and the use of the rejective method in the NPHS see Singh *et. al*[8].

3.3 Application of the Rejective Method to the NPHS

The rejective method was applied in most strata outside of Quebec. Exceptions included the LFS apartment and remote strata. In remote regions, the cost of visiting a dwelling was much higher than other

areas so it was not advisable to reject a household that costs so much to contact. Most households in the apartment strata are small. Rather than applying the rejective method in these cases, the sample size in the apartment strata was reduced.

A number of different options for the rejection rule (the characteristic that an EFR household must have in order to be rejected) were considered. Rejection rules based on household size (less than three members), members' ages (no one under 20 or 25) or a combination of both were examined. Finally, it was decided to reject households with no member under 25 years old as it required fewer rejections to achieve comparable improvements in the representativity for targeted groups. The rate at which dwellings were assigned as EFR ranged between 18.75% and 40% depending upon the location of the stratum. In rural areas, the rate was lower since the collection costs in these areas were higher. The rate was also lower in some regions of provinces where sub-provincial estimates were required. Below are the final results from the NPHS in terms of rejected households.

Number of Households in the NPHS:

Total:	20725
Total in LFS Strata Outside Quebec:	16928
Total Responding EFR:	6443 (38.1%)
Total Rejected:	3447 (20.4%)

Table 2 compares the age distribution of the NPHS selected member when the rejective method was used to what it would have been without the rejective method. For comparative purposes, the population distribution from the 1991 Census is also included.

Table 2
Age Distribution of the NPHS Selected Member: With and Without Rejective Method

Age Group	1991 Census	NPHS with rejective method	NPHS without rejective method
0-11	16.7%	11.9% (-28.5)	9.9% (-40.5)
12-24	18.2%	16.4% (-9.8)	13.7% (-25.0)
25-44	34.2%	33.0% (-3.5)	32.2% (-5.8)
45-64	20.0%	22.1% (10.7)	24.4% (21.9)
65+	11.0%	16.5% (49.9)	19.8% (80.0)

Note: The values in brackets represent the percent difference between the percentage distribution for the method in question compared to that of the Census. Percent difference = 100 × (method-Census) / Census

This table shows that the under-representation of children and youths was improved by using the rejective method. That it did not improve as much for the 0-11 year old age group was mostly due to requirements for integration of the NPHS with a national survey of children. For more information see Singh *et al*[8]. In addition, this improvement did not come at the expense of parents (generally aged 25-44). The seniors, who were greatly over-represented without the rejective method, still were, but not to such a large degree.

4. Properties of the Rejective Method

In this section the theoretical properties of the rejective method are examined. This includes applying the rejective method to a simple random sample and a two-stage design. In addition, the stability of the sample size under a two-stage design is examined.

4.1 Variance Under Simple Random Sampling

Consider a population with two domains. Let a sample selected using simple random sampling without replacement (SRSWOR) result in an over-coverage of units in domain "a" and an under-coverage in domain "b". To improve the representativity of domain "b", the rejective method is applied. A subsample of units is selected by SRSWOR to be eligible for rejection (EFR). Any domain "a" units that are in the EFR subsample are rejected. Consider the impact on variance.

Notation:

N - the population size
N_a - the population size of domain "a"
N_b - the population size of domain "b"
n - the total sample size, $n = n' + n''$
n' - the sample size of not eligible for rejection units
n'' - the sample size of eligible for rejection units
δ_j - sample indicator ($\delta_j = 1$ if unit j is in the sample, $=0$ otherwise)
δ_j' - indicator for the not eligible for rejection subsample ($\delta_j' = 1$ if unit j is not EFR, $=0$ otherwise)
y_j - the response value from unit j
Y_a - the total for y from domain "a"
Y_b - the total for y from domain "b"

An estimate of the total for a variable y under the rejective method is then $\hat{Y}_{rej} = {}_a\hat{Y}' + {}_b\hat{Y}$ where ${}_a\hat{Y}'$ and ${}_b\hat{Y}$ are estimates of Y_a and Y_b based on samples of size n' and n respectively, that is

$$ {}_a\hat{Y}' = \frac{N}{n'}\sum_{N_a} \delta_j' y_j \quad \text{and} \quad {}_b\hat{Y} = \frac{N}{n}\sum_{N_b} \delta_j y_j. $$

The variance of such an estimate can be written as

$$V(\hat{Y}_{rej}) = V({}_a\hat{Y}') + V({}_b\hat{Y}) - \frac{2(N-n)Y_aY_b}{n(N-1)},$$

where $V({}_a\hat{Y}')$ and $V({}_b\hat{Y})$ are variances of the domain estimates, that is

$$V({}_a\hat{Y}') = N^2(1-f')\,{}_aS^2$$

$$\textit{where } f' = \frac{n'}{N} \textit{ and } {}_aS^2 = \frac{(\sum_{N_a} y_j^2) - \frac{Y_a^2}{N}}{N-1},$$

and

$$V({}_b\hat{Y}) = N^2(1-f)\,{}_bS^2$$

$$\textit{where } f = \frac{n}{N} \textit{ and } {}_bS^2 = \frac{(\sum_{N_b} y_j^2) - \frac{Y_b^2}{N}}{N-1}.$$

Comparatively, the variance for a simple random sample without replacement where there is no rejective method in place can be written as

$$V(\hat{Y}_{SRS}) = \frac{N^2(1-f^*)S^2}{n^*}$$

$$\textit{where } f^* = \frac{n^*}{N}$$

$$\textit{and } S^2 = \frac{\sum_N (y_j - \bar{Y})^2}{(N-1)}.$$

Here n^* is the sample size that will produce the same overall cost as a rejective sample of initial size n with n′ units not eligible for rejection. This means that for the same cost, the difference in variance would be

$$V(\hat{Y}_{SRS}) - V(\hat{Y}_{rej}) = N^2\left[S^2\left[\frac{1}{n^*} - \frac{1}{n}\right] - {}_aS^2\left[\frac{1}{n'} - \frac{1}{n}\right]\right].$$

For the same collection costs, $n' \le n^* \le n$, since the rejected households cost much less to complete than responding households. Whether the rejective method produces lower variances than SRSWOR without rejection depends upon the variation within and the distribution of the sample between the domains.

For equal collection costs, the relationship between n^*, n′ and n can be written as

$$n^* = n - (n-n')(1-c_r)\frac{N_a}{N},$$

where c_r is the relative cost of a rejected unit ($c_r < 1$).

4.2 Variance Under a Two-stage Design

The next step in examining the properties of the rejective method is to add the one person per household selection rule. This is now a two-stage design where first the sample households are selected with SRSWOR and then within each non-rejected household, an individual is chosen at random.

Let X_j be the number of people in household j

y_{jk} be the y value for member k in household j

and δ_{jk} be an indicator that member k is in the sample ($\delta_{jk}=1$ if member k is in the sample, $=0$ otherwise)

For a variable y the estimate and within household variance for household j are

$$\hat{y}_j = X_j \sum_{X_j} \delta_{jk} y_{jk}$$

$$V_2(\hat{y}_j) = X_j^2 \sigma_j^2,$$

$$\textit{where } \sigma_j^2 = \sum_{X_j} \frac{(y_{jk} - \bar{y}_j)^2}{X_j}$$

and $\bar{y}_j$ is the average y value for household j.

The two-stage variance can be written as $V(\hat{Y}) = V_1E_2(\hat{Y}) + E_1V_2(\hat{Y})$ where V_1E_2 is the contribution to the variance from selecting a sample of households while E_1V_2 is the component describing the variance due to selecting one member in each household.

When y_j is defined as the total y value from household j, the value of $V_1E_2(\hat{Y})$ is the same as the variance from Section 4.1 while

$$E_1V_2(\hat{Y}_{SRS}) = \frac{N}{n^*}\sum_N X_j^2\sigma_j^2$$

$$E_1V_2(\hat{Y}_{rej}) = \frac{N}{n'}\sum_{N_a} X_j^2\sigma_j^2 + \frac{N}{n}\sum_{N_b} X_j^2\sigma_j^2,$$

so

$$V(\hat{Y}_{SRS}) - V(\hat{Y}_{rej}) = \left[\begin{array}{l}\left(\frac{1}{n^*} - \frac{1}{n}\right)(N^2S^2 + N\sum_N X_j^2\sigma_j^2) \\ -\left(\frac{1}{n'} - \frac{1}{n}\right)(N^2\,{}_aS^2 + N\sum_{N_a} X_j^2\sigma_j^2)\end{array}\right].$$

Once again, for the same collection costs, the variance of the rejective method may be better or worse than under SRS, depending upon the characteristics of the two domains and the distribution of the sample.

4.3 Random Sample Size

One of the disadvantages of the rejective method is that the effective sample size is not fixed, but rather dependent upon the number of dwellings that are rejected. This is not a major problem if the sample size is still relatively stable. This section gives an indication of the stability of the sample size.

A simple two-stage sample design is illustrated, where

m clusters are selected using SRSWR, and n dwellings are chosen from each cluster using SRSWOR. Assume that an EFR subsample of n" dwellings from n is identified independently within each cluster. EFR units in domain "a" are rejected.

For sample cluster i, the effective cluster sample size after rejections, n_{ei}, has second-stage expected value

$$E_2(n_{ei}) = n - n'' r_{ai} ,$$

where $r_{ai} = N_{ai} / N_i$ is the proportion of units in domain "a" for the cluster. The second-stage variance of n_{ei},

$$V_2(n_{ei}) = \frac{(N_i - n'') n'' r_{ai} (1 - r_{ai})}{(N_i - 1)} ,$$

is less than or equal to n" r_{ai} (1 - r_{ai}) given that n" is not zero.

To calculate the variance of the total effective sample size after rejections requires taking the first-stage variance and expectation, respectively, of these two equations. As the first-stage is with replacement,

$$E_1V_2(\sum_m n_{ei}) \le E_1(\sum_m n'' r_{ai} (1 - r_{ai}))$$
$$= n'' m (\bar{R}_a - \bar{R}_a^2 - \sigma_{ra}^2) ,$$
$$where \quad \bar{R}_a = \frac{1}{M} \sum_M r_{ai} ,$$
$$and \quad \sigma_{ra}^2 = \frac{1}{M} \sum_M (r_{ai} - \bar{R}_a)^2 .$$

We also have

$$V_1E_2(\sum_m n_{ei}) = V_1(\sum_m (n - n'' r_{ai}))$$
$$= n''^2 m \sigma_{ra}^2 .$$

A maximum value for $\bar{R}_a - \bar{R}_a^2$, obtained when $\bar{R}_a$ = 1/2, is 1/4. A reasonable maximum for σ_{ra}^2, obtained when the frequency distribution of the r_{ai}'s is flat over the interval (0,1), is 1/12 (the actual maximum is 1/4). Substituting these two maxima gives a reasonable maximum variance for the effective sample size of n" (n"+2) m / 12.

Table 3 below gives, for combinations of n" and m, the corresponding "maximum" standard errors for the effective sample size. For illustrative purposes, expected sample sizes are given when $\bar{R}_a$ = 1/2 and when the EFR rate is at 50% (i.e., when n = 2n"). As can be seen, the variability of the sample size is reasonably small, corresponding to CV's below 5% in all cases shown.

Table 3
Expected Sample Sizes and Standard Errors of Sample Sizes Under Different Two-stage Scenarios

		Sample Sizes			Standard Errors		
n"	n	25	100	400	25	100	400
4	8	150	600	2400	7	14	28
10	20	375	1500	6000	16	32	63
20	40	750	3000	12000	30	61	121
40	80	1500	6000	24000	59	118	237

5. Illustrative Study of the Rejective Method

In order to see the effects of changing different parameters in a rejective method sample, an empirical study was done where these parameters could be easily altered. By examining the results, it is possible to get an idea of where the rejective method is successful and where it is not.

The study population was one-third of the households that answered the 1991 Census 2B (long) form in New Brunswick. This represents a sample of approximately 1/15th of the entire provincial population. Estimates and variances were computed for several variables under different applications of the rejective method.

Two definitions of domain "a" households, corresponding to two types of rejection rules were tried. The first rule was the same as the one used for the NPHS, that is, if an EFR household had no member under 25 years of age it was rejected. The second rule rejected any EFR household with only one member. In New Brunswick, 46% of the households have no one under 25 while only 18% are one-person households. These households account for 29% and 6% of the total population respectively.

Four different values were given for c_r, the relative cost of a rejected household compared to a non-rejected one - 1/3, 1/5, 1/7, 1/9.

The eligible for rejection rates (n"/n) were also varied throughout the study. The EFR rates were set at 0% (no rejections), 20%, 40% and 60%.

Four person level variables were estimated (their prevalence in the population is shown in brackets). They included a disability indicator (9.9%), a French mother tongue indicator (33.3%), an indicator of people born outside New Brunswick (16.9%) and personal income (restricted to people aged 15 and over).

Both a two and a three-stage design were examined.

In both designs, Census Enumeration Areas (EAs) were stratified by urban/rural status, geographical proximity and average income to produce a total of 50 strata. It was then assumed that two EAs were selected from each stratum using probability proportional to size with replacement (PPSWR). Within a selected EA, it was assumed that households were selected using SRSWOR. From the basic SRS design (no rejection), $n^* = 10$ households were selected from each EA. The sample sizes (n' and n'') from an EA for a rejective design depended upon the relative cost of rejection (c_r) and the EFR rate, total costs remaining the same as those for the basic design (see section 4.1). The third stage in the three-stage design was the random selection of one person from each household.

If h represents a stratum, i is a cluster and X is the population count, then the total estimate and variance for a characteristic y under such a design is

$$\hat{Y} = \sum_h \hat{Y}_h = \sum_h \frac{X_h}{2} \sum_{i=1,2} \frac{\hat{Y}_{hi}}{X_{hi}}$$

$$V(\hat{Y}) = \sum_h V(\hat{Y}_h) = \sum_h \frac{1}{2} \left[X_h \sum_{M_h} \frac{Y_{hi}^2}{X_{hi}} - Y_h^2 \right],$$

where M_h is the number of clusters in stratum h and $\hat{Y}_{hi}$ is the estimated total for y for cluster i in stratum h.

Using components previously derived, a three-stage variance can be defined as

$$V(\hat{Y}) = V_1E_2E_3(\hat{Y}) + E_1V_2E_3(\hat{Y}) + E_1E_2V_3(\hat{Y}),$$

where the three terms on the right of the equal sign describe the contributions to the variance from selecting clusters, households and individuals within households respectively.

Table 4 shows the impact that the rejective method has on both the initial and effective sample sizes under some of the combinations of parameters described above.

As households with no member under 25 are more frequent that one-member households, applying a rejection rule based on the former has a bigger impact on the effective sample size. Tripling the cost ratio c_r from 1/9 to 1/3 has a relatively minor impact on the sample sizes for these cases.

Table 4
Impact of the Rejective Method on Sample Size

Rejection Rule = = >			No one under 25		1 person household	
Cost (c_r)	Sample Size	EFR rate 0%	20%	60%	20%	60%
1/3	Total	1000	1066	1227	1025	1079
	Effective	1000	967	887	988	960
1/9	Total	1000	1090	1327	1034	1109
	Effective	1000	989	959	996	986

Table 5 shows the influence of the rejective method on the distribution of the sample by age. Both two and three-stage designs are examined. Values given for each age group and strategy are the differences between the sample sizes and the sample sizes that would have been obtained under a two-stage design with no rejection, expressed as a percentage of the latter.

Table 5
Impact of the Rejective Method on Sample Distribution
Rejection Rule: No one under 25, $c_r = 1/9$

Age Group	Two-stage design 20%	60%	Three-stage design 0%	20%	60%
0-14	3.4	18.3	-29.6	-24.7	-5.4
15-24	6.7	22.2	-15.7	-6.1	18.2
25-44	0.8	2.0	-4.6	-3.0	-1.0
45-64	-3.4	-16.7	18.3	13.5	-3.9
65+	-12.5	-43.2	61.7	42.5	-4.4

A two-stage design with no rejective method applied has a population distribution similar to the Census distribution since no households are being rejected and all members are selected. When the rejective method is applied in such a design, the result is that many of the dwellings that contain older people are rejected. This means that the sample is losing some of its older members but keeping all of its younger ones. Thus the sample has an over-representation of young people and an under-representation of older members. As the EFR rate increases, this mis-representation grows. On the other hand, in the three-stage design only a single member of the household is selected. This results in an under-representation of younger people when there is no rejective method in place (see Table 1 for another example). As the EFR rate increases, a greater percentage of the retained dwellings contain younger people, so the distribution of the selected members moves closer to the distribution of the entire population. This improvement under the three-stage design is obvious

from the table. In fact the 60% EFR column indicates a situation where some of the age categories that were over-represented are now under-represented and vice versa.

Table 6 shows some of the results of changing the design, rejection rule, and rejection rates on the coefficients of variation (CVs) of the variables in question at both the overall and domain levels. The cost ratio used is 1/5.

At the overall level there are small impacts on the CVs, and in most cases they rise as the eligible for rejection rate increases. However there are some exceptions to this rule. When the domain "a" households are one person households, there is a decrease in CVs under the three-stage design for the born outside New Brunswick and French mother tongue variables. As well, when using the no one under 25 rejection rule and a three-stage design, the born outside New Brunswick and income variables actually have a decrease in CV for the 20% and 40% EFR rates although it is negligible.

At the overall level the effects of increasing the EFR rate has a larger impact on the results for the two-stage design than the three-stage design. There is also a greater effect when the domain "a" dwellings are more prevalent in the population as a whole. This is shown by the larger changes that take place when the no one under 25 rule is used compared to the one person per household rule. Although space does not permit us to show the results, another characteristic is that the changes in CVs are larger when a smaller cost ratio (c_r = 1/9) is used.

As expected, the changes at the domain level are more pronounced. The CVs for domain "a" (dwellings which fail the screening criterion) increase while those for domain "b" decrease as the rejection rate increases. This is not surprising since the rejective method allocates more dwellings to domain "b" and fewer to domain "a" than a non-rejective method does. The impact on the CVs at the domain level is sometimes fairly large and most often the domain "a" results suffer to a larger degree than the domain "b" CVs improve.

Figure 2 shows the impact of the rejective method on the CVs of different age categories. For each characteristic there are five points plotted (except income which only has four). Each point represents an age category (from left to right 0-14, 15-24, 25-44, 45-64 and 65+). There is no 0-14 group for income. Each symbol represents a different design.

Table 6
CVs by Domain under Different Scenarios (c_r = 1/5)

6a: Rejection Rule: No one under 25 years of age

	2-stage design				3-stage design			
EFR rate = = >	0%	20%	40%	60%	0%	20%	40%	60%
Domain "a": Households with nobody under 25								
disabled	13.4	13.8	14.5	15.8	14.5	15.1	16.0	17.6
french	22.8	23.0	23.3	23.9	22.9	23.1	23.4	24.1
born out NB	21.8	22.1	22.6	23.6	22.6	23.0	23.7	25.0
income	7.9	8.1	8.6	9.4	8.6	8.9	9.5	10.5
Domain "b": Households with persons under 25								
disabled	21.4	21.1	20.9	20.6	26.3	25.8	25.2	24.6
french	10.2	10.1	10.0	9.9	10.4	10.3	10.2	10.1
born out NB	11.8	11.6	11.4	11.2	13.7	13.4	13.1	12.8
income	5.5	5.4	5.3	5.1	8.5	8.2	8.0	7.7
Overall: All households								
disabled	9.3	9.5	9.8	10.5	11.4	11.6	11.9	12.7
french	7.0	7.0	7.1	7.2	7.2	7.2	7.2	7.3
born out NB	8.2	8.2	8.3	8.5	9.7	9.7	9.7	9.8
income	3.3	3.4	3.5	3.8	5.3	5.3	5.3	5.5
Sample Sizes:								
Total	1000	1080	1174	1285				
Not EFR	1000	864	704	514				
Effective	1000	980	957	929				

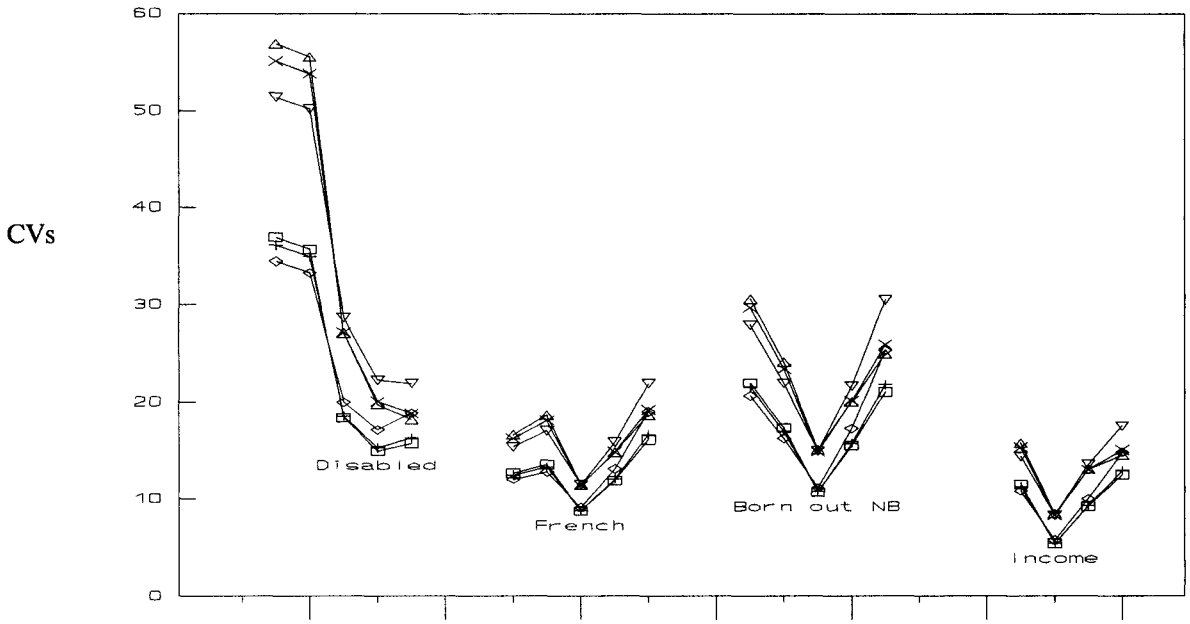
6b: Rejection Rule: One person households

	2-stage design				3-stage design			
EFR rate = = >	0%	20%	40%	60%	0%	20%	40%	60%
Domain "a": Households with only one member								
disabled	45.7	46.3	47.2	48.9	45.7	46.3	47.2	48.9
french	106.	106.	106.	107.	106.	106.	106.	107.
born out NB	83.8	84.1	84.7	85.8	83.8	84.1	84.7	85.8
income	27.3	27.7	28.5	29.9	27.3	27.7	28.5	29.9
Domain "b": Households with more than one member								
disabled	10.5	10.5	10.4	10.3	13.2	13.0	12.9	12.8
french	7.5	7.5	7.4	7.4	7.7	7.6	7.6	7.6
born out NB	8.7	8.7	8.6	8.6	10.4	10.3	10.2	10.1
income	3.7	3.6	3.6	3.6	5.9	5.8	5.7	5.7
Overall: All households								
disabled	9.3	9.3	9.3	9.5	11.4	11.4	11.4	11.4
french	7.0	7.0	7.0	7.0	7.2	7.2	7.1	7.1
born out NB	8.2	8.2	8.2	8.2	9.7	9.7	9.6	9.6
income	3.3	3.3	3.3	3.4	5.3	5.3	5.3	5.3
Sample Sizes:								
Total	1000	1030	1063	1097				
Not EFR	1000	824	638	438				
Effective	1000	992	984	976				

The first plot represents the rejection rule of no member under 25. Here the overall CVs decrease for

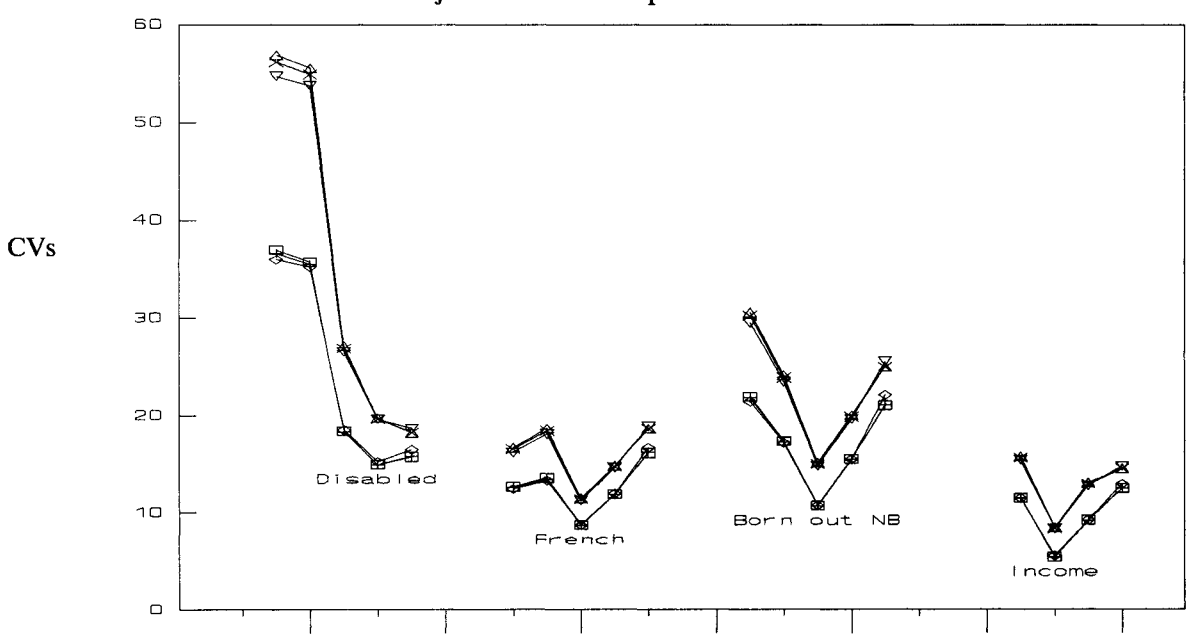
Figure 2
Resulting CVs from Different Scenarios (c_r=1/5)

Rejection Rule: No one under 25 years of age



Characteristics and age-groups

Rejection Rule: One person Households



Characteristics and age-groups

□ - two-stage, 0% efr
+ - two-stage, 20% efr
◇ - two-stage, 60% efr
△ - three-stage, 0% efr
× - three-stage, 20% efr
▽ - three-stage, 60% efr

younger people and increase for older members as the EFR rate goes up. This change is very minor, and is most noticeable for the disabled and born outside of New Brunswick variables. The drop is also consistent with the redistribution of sample to households with younger people. Generally, the increase in CVs for seniors is larger than the decrease for young people.

Although the same general results occur, the effect is less obvious when the rejection rule is one member households. This is because there are fewer households in this category in the population so the redistribution of the sample is less pronounced.

The impacts on CVs by age are very minor. The advantages of a rejective method should be examined very carefully before applying it. In fact it may be used more to improve the sample representativity of certain sub-populations, or to ensure adequate sample yields for them, rather than to improve their CV estimates.

6. Conclusions

In this report an attempt has been made to illustrate the properties of the rejective method used in the NPHS. The illustrative examples given show a basic application of the method. Extra layers of complexity could be added by changing EFR rates between strata, requiring integral sample sizes, etc.

The driving force behind using the rejective method for the NPHS was to improve the representativity of the sample which is affected by the use of the one person per household rule. In this sense the rejective method was successful, since more young people were added to the sample at the expense of seniors who were initially over-represented.

The redistribution of sample brings it closer to the population's distribution, therefore it was hoped that the overall variances would decrease as well. In this sense the results were not as encouraging. Overall, the CVs generally suffered slightly when the rejective method was applied compared to a usual non-rejective method. As the percentage of dwellings eligible for rejection increased, this rise in overall CVs became more obvious.

At the domain level, the results were as expected. The variance of estimates within domain "b" was reduced. On the other hand, the variances in domain "a" (where dwellings were eligible for rejection) increased. This was anticipated since a greater proportion of the sample was in domain "b" under the rejective method. A similar situation occurred at the age group level where the variances for children and youths dropped while those for older people increased as the eligible for rejection rate rose.

A positive characteristic was the stability of the sample size. Although there is a degree of randomness in the size of the sample due to the implementation of the rejective method, the variability is reasonably small in comparison to the overall targeted sample size. As long as the user has a good idea of the percentage of units in the population which fall into domains "a" and "b", the effective sample size can be confidently estimated.

Before using a rejective method, the user should examine different components and goals of his/her survey to determine if the rejective method is appropriate.

References

[1] Stephens T., and Fowler G.D., editors (1993). *Canada's Health Promotion Survey 1990: Technical Report*. Catalogue No. H39-263/2-1990E, Health and Welfare Canada.

[2] Norris D.A. and Paton D.G. (1991). Canada's General Social Survey: Five Years of Experience. *Survey Methodology* 17, 227-240.

[3] D. Brown. (1994). The 1992-93 New Zealand Household Health Survey. *The Survey Statistician* 30, 10-12.

[4] Statistics Canada and Health and Welfare Canada (1981). *The Health of Canadians: Report of the Canada Health Survey*. Statistics Canada, Catalogue 82-538E, Statistics Canada.

[5] Ontario Ministry of Health (1982). *Ontario Health Survey Highlights*.

[6] Courtemanche R. and Tarte F. (1987). *Sampling Plan for the Quebec Health Survey. Technical Manual.* Enquête Santé Québec, 87-02, Santé Québec.

[7] Adams P.F. and Benson V. (1991). Current Estimates from the National Health Interview Survey. *Vital Health Statistics* 10,181, National Center for Health Statistics.

[8] Singh M.P., Tambay J.L., Krawchuk S. (1994). The National Population Health Survey: Design and Issues. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 803-808.