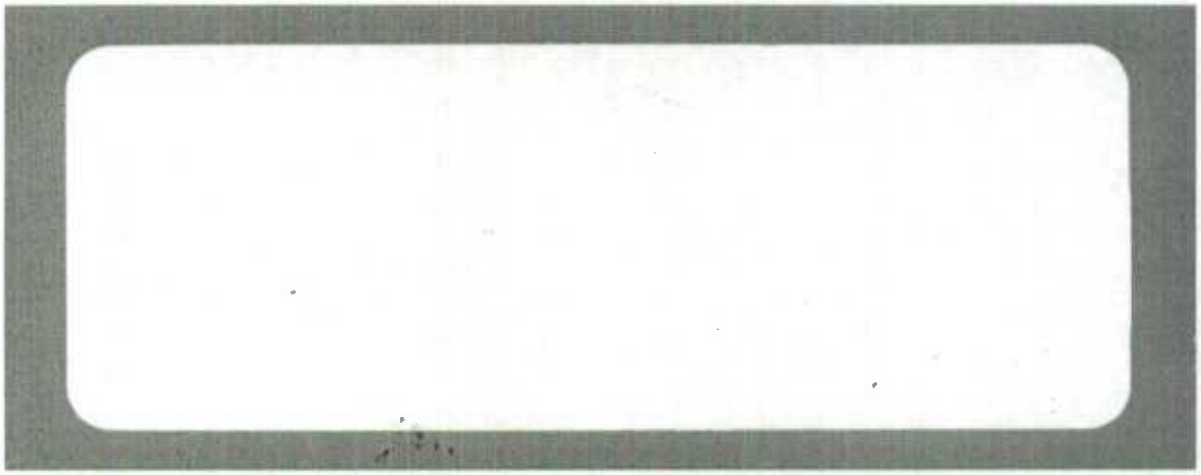
Methodology Branch

Direction de la méthodologie



Household Survey Methods Division

Division des méthodes d'enquêtes auprès des ménages

Statistics Canada Statistique Canada

Canada

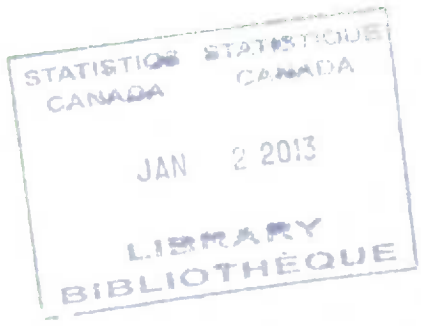
WORKING PAPER
METHODOLOGY BRANCH

EVALUATING THE DESIGN EFFECT AT DIFFERENT STEPS OF THE WEIGHTING PROCESS IN THE CANADIAN COMMUNITY HEALTH SURVEY

HSMD - 2012 – 006E
Golshid Chatrchi, Marie-Claude Duval, Francois Brisebois, Steven Thomas

Household Survey Methods Division
Statistics Canada

October 2012



Evaluating the Design Effect at Different Steps of the Weighting Process in the Canadian Community Health Survey

Golshid Chatrchi, Marie-Claude Duval, Francois Brisebois, Steven Thomas

ABSTRACT

The Canadian Community Health Survey (CCHS) is a cross-sectional survey that collects general health information on the Canadian population at the Health Region (HR) level. The survey will be redesigned in 2015. One component of the redesign is the improvement of the efficiency of the sample design. The CCHS uses a complex survey design which includes multiple frames and multiple stages of selection within each of those frames. Weights resulting from the design are adjusted through several steps such as frame integration, nonresponse and calibration. The combination of these complexities leads to the production of estimates with relatively high variances when compared to what would be obtained through a simple random sample. Design effects are examined in this paper to quantify how the overall variability in estimates can be attributed to the complex survey design and to each of the individual adjustments in the weighting process. The findings will help guide possible modifications for the upcoming redesign.

Key words: design effect, weighting, complex survey design, multiple frames, bootstrap

Évaluation des effets de plan aux différentes étapes de la pondération dans l'Enquête sur la santé dans les collectivités canadiennes

RÉSUMÉ

L'Enquête sur la Santé dans les collectivités canadiennes (ESCC) est une enquête transversale qui recueille de l'information sur la santé générale de la population canadienne au niveau des régions sociosanitaires. L'enquête sera remaniée en 2015. Une des composantes du remaniement est l'amélioration de l'efficacité du plan d'échantillonnage. L'ESCC utilise un plan de sondage complexe qui inclut l'utilisation de bases multiples et la sélection d'un échantillon sous un plan complexe à plusieurs degrés. Les poids de sondage initiaux sont ajustés à différentes étapes telles que l'intégration des bases, la non-réponse et le calage. La combinaison de ces ajustements produit des estimations avec variances relativement élevées comparées aux variances sous un plan aléatoire simple. Dans cet article, les effets de plan sont examinés afin de quantifier la variabilité dans les estimations sous ce plan complexe ainsi que pour chaque ajustement considéré dans la pondération. Les résultats seront utiles dans l'élaboration du nouveau plan de sondage du prochain remaniement.

Mots clés : effet de plan, pondération, plan de sondage complexe, bases de sondage multiples, bootstrap.

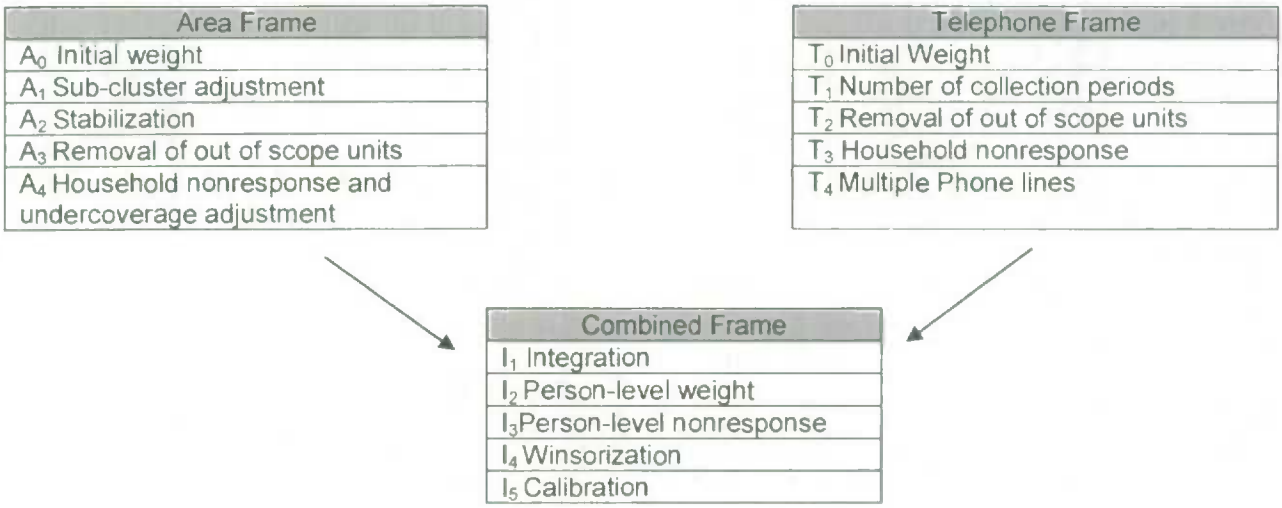
1. INTRODUCTION

The CCHS is a cross-sectional survey that collects information related to health status, health care utilization and health determinants for the Canadian population. It relies upon a large sample of respondents and is designed to provide reliable estimates at the health region level (HR). In 2007, major changes were made to the survey design with the goal of improving its effectiveness and flexibility through a continuous collection approach. Data collection now occurs on a continuous basis with data products released every year, rather than every two years as was the case prior to 2007. To provide reliable HR-level estimates, a sample of 65,000 respondents is selected annually. A multi-stage sample allocation strategy gives relatively equal importance to the HRs and the provinces. In the first step, a minimum of 250 units is imposed on each HR annually. The second step is to allocate the remaining sample to the provinces proportionally to the size of their population. The third step is to allocate each province's additional sample to the HRs proportionally to the square root of their population. Finally, for some HRs, another level is added to the stratification to answer more precise geographic needs at the sub-HR level. The HR sample is allocated to the sub-HRs proportionally to their population size.

The 2010 CCHS design used three sampling frames to select the sample of households: 49.5% of the sample of households came from an area frame, 49.5% came from a list frame of telephone numbers and the remaining 1% came from a Random Digit Dialing (RDD) sampling frame. The area frame used by the CCHS is the one designed for the Canadian Labor Force Survey (LFS). The LFS design is a multi-stage stratified cluster design in which the dwelling is the final sampling unit. 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 are selected from these lists. The CCHS basically follows the LFS sampling design with health regions generally forming the strata. The telephone list frame is the InfoDirect list, an external administrative database of names, addresses and telephone numbers from telephone directories in Canada. Within each stratum (generally the HR), the required number of telephone numbers is selected using a Simple Random Sampling (SRS) design. For both frames, once a contact is made with a household, one person is selected from all household members based on unequal person selection probabilities.

Given this complex survey design, several steps of weighting adjustments are required. The weighting steps can be summarized in diagram 1. For more details about the below diagram and weighting process, refer to the CCHS user guide (Statistics Canada (2011)).

Diagram 1: Weighting process



The goal of this study is to estimate the variability caused by the sampling design and each step of the weighting process. In order to do this, design effects have been computed for each of these steps using the 2010 CCHS data. The design effect is the ratio of the sampling variance of an estimator under a given design to the sampling variance of an estimator under a Simple Random Sampling (SRS) design with the same sample size. This is a measure used to evaluate the efficiency of a design. If the ratio is less than one, this indicates that the sample design is more efficient than a SRS; greater than one indicates that the sample design is less efficient than a SRS. The design effect has an impact on the sample size required to do an analysis. The larger the design effect, the more sample required to obtain the same precision of an estimate as would have been obtained under an SRS design. This is why small design effects are desired when developing a sampling design. For more information on design effects, refer to Kish (1965).

2. METHODOLOGY OF THE STUDY

Since the early steps of the weighting process are performed at the household level, variables that relate to household characteristics were used throughout this study. Also, since there is a potential relation between the design effect and the estimate of a proportion, variables with different prevalence were chosen to get more exhaustive results. The four variables chosen below have prevalence rates between 5% and 20% at the Canada level. For the variable definitions and specifications, see Appendix A.

The four variables are:

- Households with at least one child~ 20%
- Households with one or two kids~10%
- Households with five members~5%
- Households with exposure to second-hand smoke~10%

The design effects of these variables are calculated after each of the following weighting adjustments. The codes are based on Diagram 1:

(1) Removal of out-of-scope units- A_3 and T_2,
(2) Household nonresponse- A_4 and T_3,
(3) Integration- I_1,
(4) Person-level weight- I_2,
(5) Person-level nonresponse- I_3,
(6) Winsorization- I_4 and
(7) Calibration- I_5.

The design effects are calculated for each of the four variables at the health region level, which is the main domain of interest. They are also calculated at the provincial level in order to evaluate the design effect at that level knowing the sample allocation is not optimal. Finally it is calculated at the CCHS stratum level, which should show the real impact of the CCHS design without the effect of the sample allocation method.

3. ANALYSIS

Since the RDD frame only covers 4 HRs, which represent a small portion of the target population, the RDD sample was not included in this study. As well, the Yukon, Northwest Territories and Nunavut are excluded. The sampling method used for the three territories is slightly different than used in the

provinces, which would have added a lot of complexities to the comparisons. Also, only the estimates with at least 5 observations in the numerator before integration and 10 observations in the numerator after combining the two frames were kept in the study to ensure good data quality.

As mentioned earlier, the design effects were calculated for each variable and for each domain of interest (Health Region, Province, CCHS Stratum) after each weighting step. To obtain the design effects, variances were calculated with the BOOTVAR software using the bootstrap weights from each weighting step, with the use of the BOOTVAR DEFF option to get the design effects. BOOTVAR is a statistical product which was created to estimate variances using the bootstrap method and includes options for many types of estimates, including proportions. It is available in SAS and SPSS programming languages (Statistics Canada (2011)). In the CCHS, the bootstrap method used is the rescaling bootstrap method proposed by Rao and Wu (1988).

Section 3.1 looks specifically at each of the 7 weighting steps listed in section 2. Results are focused on the HR level. Results at the province and CCHS stratum levels can be found in Appendix C. Section 3.2 provides a graphical summary of the weighting steps for the three domains examined (province, HR and CCHS stratum).

3.1 Design effects for each weighting step

In each sub-section below, a short description of the weight adjustment is given followed by a summary of the design effects at the particular step for the four variables of interest. More details about these steps can be obtained in the CCHS user guide (Statistics Canada (2011)).

3.1.1 Removal of out-of-scope units (Steps A_3 and T_2)

The first step in evaluating the CCHS process was to closely examine the effect of the sampling design on the area and telephone frames. The weight obtained just after the out-of-scope removal was chosen for this evaluation partly because of practical reasons.

During collection, some sampled dwellings are found to be out-of-scope. For example dwellings that are demolished, under construction or institutions are identified as being out-of-scope in the area frame while numbers that are out of order or for business are out-of-scope on the telephone frame. These units and their associated weights are removed from sample.

At this step we are interested in calculating estimates that include both responding and nonresponding households. However, at this point in the process the CCHS data files would only contain information on the variables of interest from responding households. Therefore, data for the nonresponding households had to be imputed at this step. The details of the imputation method can be found in Appendix B. Once applied, each household had a valid response for each of the four variables of interest. Table 1 presents the design effect distribution at the HR level, after removal of out of scope units.

Table 1: Summary of the design effect distribution for the four variables at the HR level after removal of out of scope units

Variable	Frame	Min	P 25	P 50	P 75	Max	Mean
% of households with at least one child (20%)	Area	0.46	1.08	1.26	1.60	4.64	1.38
	Tel	0.87	0.99	1.03	1.07	1.33	1.04
% of households with one or two kids (10%)	Area	0.29	1.04	1.23	1.55	2.95	1.33
	Tel	0.86	0.96	1.02	1.07	1.37	1.02
% of households with five members (5%)	Area	0.23	0.97	1.20	1.43	3.96	1.26
	Tel	0.85	0.95	1.00	1.04	1.51	1.01
% of households with exposure to second-hand smoke (10%)	Area	0.55	1.00	1.18	1.43	3.49	1.28
	Tel	0.84	0.97	1.02	1.06	1.20	1.02

The median of health region design effects for each variable fluctuates around 1.2 for the area frame and around 1 for the telephone frame. These results are expected since the telephone frame sampling design is a simple random sampling process within each stratum while the sampling design of the area frame is a multi-stage stratified cluster design.

3.1.2 Household Nonresponse (Steps A_4 and T_3)

The second stage of the weighting process that is considered is household nonresponse. Household nonresponse occurs when a household refuses to participate in the survey or cannot be reached for an interview. These units are removed from the sample and their weights are redistributed to responding households within response homogeneity groups. These groups are created based on logistic regression models that divide the sample into groups with similar response properties. The following adjustment factors are calculated within each response group:

Area Frame:

$$\frac{sum\ of\ weight\ after\ step\ A3\ for\ all\ households}{Sum\ of\ weight\ after\ step\ A3\ for\ all\ responding\ households}$$

Telephone Frame:

$$\frac{sum\ of\ weight\ after\ step\ T2\ for\ all\ households}{Sum\ of\ weight\ after\ step\ T2\ for\ all\ responding\ households}$$

The weights after the steps A3 and T2 of responding households are multiplied by these factors to produce the weights A4 and T3.

To evaluate the effect of the household nonresponse adjustment, we computed the design effect just before the integration step. It should be noted that before the integration step, we have the effect of "Multiple Phone lines- T4" adjustment in the telephone frame which has a small impact on the design effect since it affects very few units. This step was not considered in this study. An adjustment in the area frame is also done before the integration to account for the under-coverage. The area frame has about 12 % households' under-coverage in the current LFS design. In order to deal with this frame defect,

a post-stratification adjustment is applied at the HR level using the most up-to-date household counts. This adjustment has no impact on the design effect at the HR level since it does not affect the estimates calculated in each replicate at this level and has a very small impact at the provincial level. Table 2 presents a summary of the design effect distribution for the four variables at the HR level, after the household nonresponse adjustment (including the undercoverage adjustment in the area frame).

Table 2: Summary of the design effect distribution for the four variables at the HR level after the Household nonresponse adjustment (including the undercoverage adjustment in the area frame)

Variable	Frame	Min	P 25	P 50	P 75	Max	Mean
% of households with at least one child (20%)	Area	0.41	1.20	1.43	1.73	3.72	1.52
	Tel	0.92	1.09	1.15	1.23	1.50	1.16
% of households with one or two kids (10%)	Area	0.43	1.07	1.40	1.67	3.07	1.44
	Tel	0.9	1.07	1.16	1.21	1.76	1.15
% of households with five members (5%)	Area	0.55	1.00	1.17	1.47	3.30	1.29
	Tel	0.81	1.02	1.10	1.18	1.44	1.10
% of households with exposure to second-hand smoke (10%)	Area	0.48	1.07	1.27	1.60	3.13	1.39
	Tel	0.79	1.04	1.13	1.21	1.50	1.13

Design effects increased slightly for both frames when compared to the results in Table 1. The median of the design effects at the HR level varies between 1.2 and 1.4 for the area frame and between 1.1 and 1.2 for the telephone frame. This represents an average 8% increase on the area frame and 12% increase on the telephone frame. This rise is expected since the nonresponse adjustment increases the variability of weights and therefore the variance of estimates in favor of reducing potential nonresponse bias. Also, the differences between the increases seen with the nonresponse adjustment on the two frames might be explained partly by the fact that there is more nonresponse on the telephone frame.

3.1.3 Integration (Step I_1)

The third stage of interest for the study is the impact of the integration adjustment. At this step, the two frames are merged and the design effects of the variables after the integration-I1 adjustment are calculated.

The current integration approach takes into account the portion of the population covered by both frames as well as the under-coverage of the telephone frame. Those households without a landline or without a listed telephone number are not covered by the telephone list frame. Not taking this into account could cause a bias if the households not covered by the telephone frame have different characteristics than the ones covered. To take the under-coverage into consideration, the CCHS integrates only the sampled households that are common to both frames. The weights of the households that only belong to the area frame (households without landline or without listed phone numbers) remain unchanged. This allows these households to represent other similar households in the population. For the common portion, the integration process applies a contribution to each frame. An adjustment factor α between 0 and 1 is applied to the weights; the weights of the area frame units that have telephone are multiplied by α and the weights of the telephone frame units are multiplied by $1-\alpha$. The term α represents the overall sample size contribution of the area frame to the common portion.

A composite estimator for a total thus takes the form:

$$\hat{Y}_{\text{int}} = \hat{Y}_{A}^{S_A} + \alpha \hat{Y}_{AB}^{S_A} + (1-\alpha)\hat{Y}_{AB}^{S_B}$$

Where $\hat{Y}_{int}$ represents the estimate, S_A represents the area frame sample and S_B represents the sample from the telephone frame. *AB* represents the common portion of both frames and *A* represents the portion of the area frame not covered by the list frame.

The term α has been fixed to 0.4 since 2008 for all domains which represent the overall sample size contribution of the area frame to the common portion in CCHS 2008. The reason of using a fixed α for all HRs and throughout time was to improve coherence and comparability between estimates over the domains and over time. For more information, refer to Wilder&Thomas (2010). Table 3 presents the design effect distribution at the HR level:

Table 3: Summary of the design effect distribution for the four variables at the HR level after integration

Variable	Min	P 25	P 50	P 75	Max	Mean
% of households with at least one child (20%)	0.82	1.29	1.58	1.99	4.35	1.70
% of households with one or two kids (10%)	0.90	1.29	1.57	2.00	3.94	1.68
% of households with five members (5%)	0.74	1.03	1.26	1.66	5.33	1.40
% of households with exposure to second-hand smoke (10%)	0.78	1.26	1.55	1.92	4.00	1.61

As can be seen, there is a significant jump in design effects after joining the two frames. The median design effect within the health regions was between 1.2 and 1.4 for the area frame and between 1.1 and 1.2 for the telephone frame before integration. This amount increased to around 1.3 to 1.6 after integration for the combined sample. One explanation for this change is the fact that a fixed α is used in integration rather than a more optimum one based on the coverage of the telephone frame, the sample distribution by frame and the design effect by frame for a given survey occasion and a given domain. The efficiency of the integration deteriorates as these conditions deviate from those that were observable at the Canada level when the value of α was originally determined in 2008. At that time, there was a 50 / 50 split of the sample between the area frame and the telephone frame with the telephone frame coverage of 63% of dwellings at the national level. Not all health regions have the same telephone frame coverage or the same final sample distribution by frame. As well, beginning in the fall of 2010, the sample allocation moved to a 41 / 59 sample split between the area frame and the telephone frame respectively in order to reduce collection costs. Also, in 2010, the telephone frame had an improved coverage of 66.5%. With all these differences, α is not always optimum. However, as mentioned previously, a fixed α for all HRs and throughout time was desired for consistency and comparability between the estimates over the domains and over time.

3.1.4 The Person-Level Weight (Step I_2)

The fourth stage to be studied is the derivation of the person-level weight. At this stage, the concept of the variables examined in the study is changed since the statistical unit has now become the person rather than the household. The same concepts were preserved but estimates will now reflect characteristics in terms of people instead of households.

These are:

- People living in households with at least one child (less than 12 years old)
- People living in households with one or two kids (less than 6 years old)
- People living in households with five members
- People living in households with exposure to second-hand smoke

At this step, the household-level weights are adjusted using the inverse of the person-level selection probabilities to calculate person-level weights. The person-level selection in the CCHS is done with unequal probabilities of selection based on the household size and the age of the household members. For more information on the person-level selection probabilities, refer to the CCHS user guide, Statistics Canada (2011). Table 4 presents the design effect distribution at the HR level after the creation of the person-level weight.

Table 4: Summary of the design effect distribution for the four variables at the HR level after creation of the Person-level weight

Variable	Min	P 25	P 50	P 75	Max	Mean
% of people living in households with at least one child (20%)	1.03	1.78	2.20	2.73	4.93	2.32
% of people living in households with one or two kids (10%)	1.03	1.61	1.94	2.50	4.55	2.11
% of people living in households with five members (5%)	1.09	2.00	2.39	3.21	6.56	2.68
% of people living in households with exposure to second-hand smoke (10%)	0.8	1.72	2.16	2.85	9.85	2.65

The median of the design effect within health regions varies between 1.9 and 2.4. On average this is a 55% increase compared to the design effect calculated with the integrated household weight. This rise in the design effect is mainly due to the unequal probability of selection used by the CCHS. The person-level selection adjustment can be as low as 1 and as high as 20 in some cases. This design is quite different from what would have been observed with an SRS. One other reason for the increase in the design effect could be related to the change in variable concepts from household to person level.

3.1.5 Person-level nonresponse (step I_3)

The next step of the weighting adjustments to examine is the person-level nonresponse adjustment. The CCHS interview has two parts: first the interviewer completes the list of the household's members and then one person is selected for the interview. It is possible that the household roster is obtained (household response) but the selected person refuses to be interviewed or cannot be reached for some reasons. This causes person-level nonresponse. The same treatment we used in household nonresponse is used at this stage where response homogeneity groups are created based on a logistic regression score function.

After creating response homogeneity groups, the following adjustment factor is calculated within each group:

$$\frac{sum\ of\ weight\ after\ step\ I2\ for\ all\ selected\ persons}{Sum\ of\ weight\ after\ step\ I2\ for\ all\ responding\ selected\ persons}$$

The weights after step I_2 are multiplied by this factor to create the weights I_3. Table 5 presents the design effect distribution for the four variables at the HR level, after the person-level nonresponse adjustment.

Table 5: Summary of design effect distribution for the four variables at the HR level after person-level nonresponse

Variable	Min	P 25	P 50	P 75	Max	Mean
% of people living in households with at least one child (20%)	1.21	2.02	2.43	3.10	6.09	2.67
% of people living in households with one or two kids (10%)	1.07	1.69	2.10	2.76	7.91	2.34
% of people living in households with five members (5%)	1.32	2.21	2.84	3.65	7.14	3.05
% of people living in households with exposure to second-hand smoke (10%)	0.89	1.87	2.50	3.36	17.66	2.98

The median design effect at the health region level fluctuates between 2.1 and 2.8 while the median of the design effect in the previous step was between 1.9 and 2.4. This rise is expected and represents an average 9% increase. This is similar in nature to the household nonresponse adjustment. The nonresponse adjustment increases the variability of the weights, which leads to higher variance estimates in favor of reducing potential nonresponse bias.

At this level, the maximum design effect over the health region estimates can be as large as 17.66. This is due to the presence of extreme weights which are later adjusted by Winsorization (next step).

3.1.6 Winsorization (step I_4)

The sixth stage to be examined is Winsorization. The weighting process may cause some units to have extreme weights which can have a large impact on the variance. The weights of these units are adjusted downward using a "Winsorization" trimming approach. Table 6 presents the design effect distribution for the four variables at the HR level, after Winsorization.

Table 6: Summary of design effect distribution for the four variables at the HR level after Winsorization

Variable	Min	P 25	P 50	P 75	Max	Mean
% of people living in households with at least one child (20%)	1.21	1.96	2.40	3.05	6.06	2.60
% of people living in households with one or two kids (10%)	1.07	1.66	2.04	2.68	7.91	2.27
% of people living in households with five members (5%)	1.32	2.17	2.78	3.46	6.05	2.95
% of people living in households with exposure to second-hand smoke (10%)	0.89	1.81	2.38	3.27	9.61	2.82

After the winsorization step, the median design effect varies between 2 and 2.8. Comparing this with the results of the previous step, it can be concluded that winsorization does not have a significant effect on the median of the design effect mainly because very few units are winsorized. It does deflate the extreme values that were observed for some domains in the previous step which is the goal of this process.

3.1.7 Calibration (Step I_5)

The last step to study is calibration. This adjustment is done to make sure that the sum of the final weights corresponds to the population estimates defined at the HR level for each of the 10 age-sex groups. The five age groups are 12-19, 20-29, 30-44, 45-64 and 65+, for both males and females. At the same time, weights are adjusted to ensure that each collection period is equally represented in the population. Calibration is done using CALMAR (Sautory (2003)) with the most up to date population counts and the most up to date geography boundaries. The following table demonstrates the design effect distribution for the four variables at the HR level, after the calibration adjustment:

Table 7: Summary of design effect distribution for the four variables at the HR level, after Calibration

Variable	Min	P 25	P 50	P 75	Max	Mean
% of people living in households with at least one child (20%)	0.81	1.45	1.76	2.28	3.66	1.87
% of people living in households with one or two kids (10%)	0.94	1.47	1.81	2.20	4.16	1.90
% of people living in households with five members (5%)	1.13	1.92	2.38	3.05	4.73	2.57
% of people living in households with exposure to second-hand smoke (10%)	1.18	1.76	2.26	2.69	5.97	2.40

As expected, this step significantly decreases the design effect by 14% on average for the 4 variables. The median of the design effect after calibration is about 1.8 to 2.4, compared to 2 to 2.8 at the previous step.

3.2 Domains & Variables

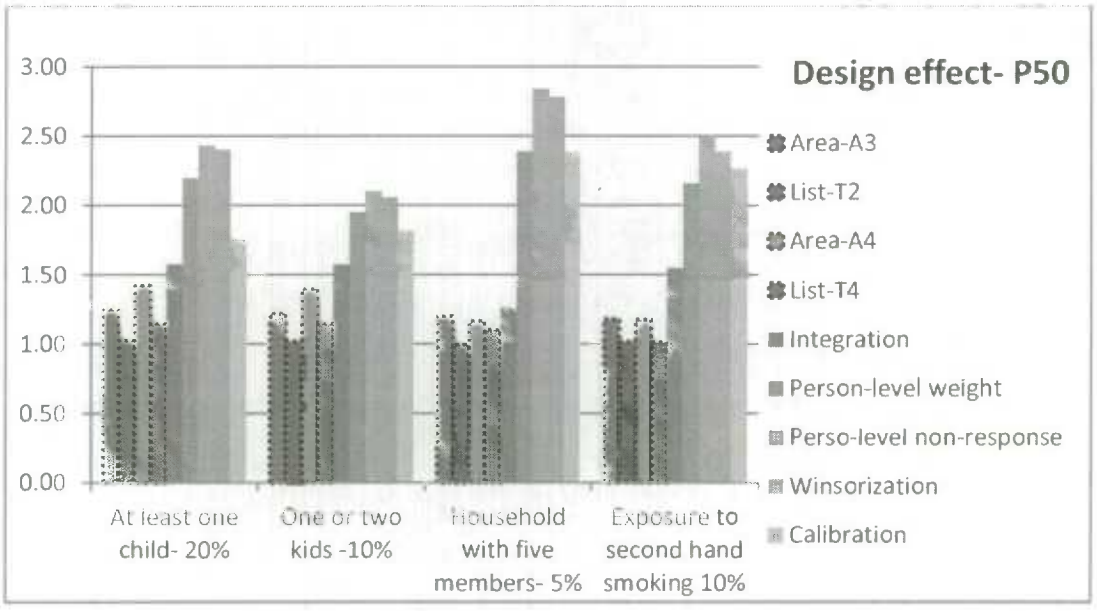
In this section, an overview of the impact each step has on the design effect is shown graphically for each of the four variables at the HR level. Graphs are also provided at the provincial and the stratum CCHS level. The graphs show the median (P50) of design effects for each step and each variable. The specific values and more details can be found in Appendix C.

3.2.1 Region level

The goal of the CCHS is to collect general health information at the Health Region (HR) level. Therefore, it is worthwhile to evaluate the changes in terms of health regions. For the 2010 CCHS, data was collected in 117 HRs, but since the Territories and two regions that only use the RDD frame were excluded, only the results from 112 HRs are used in this study.

The following chart shows the median (P50) of design effects for four variables of interest at the health region level.

Chart 1: Median (P50) of design effects at the health region level



The first four bars (dashed outlined) represent the design effects of step 1 and step 2 for the area frame and the list frame separately (before integration) and the other bars show the design effects after integration (combined frame).

The chart shows clearly that the person-level-weight and the nonresponse adjustments have the most negative impact on the design effect while calibration and winsorization decrease the design effect. The integration step has also a negative impact that is non-negligible.

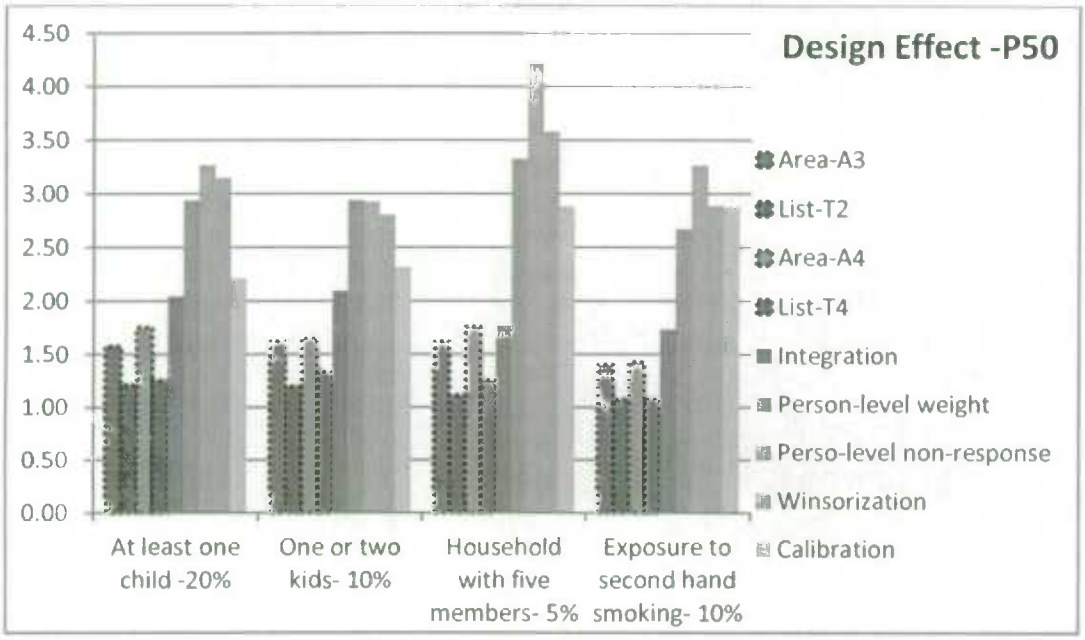
The trend of the design effects throughout the weighting process is similar for each variable. However, we can notice a larger negative impact at the person level weight for the variables 'household with five members' and 'exposure to second hand smoke' compared to the others. This is most likely due to the variability of the weights at the person level. When the prevalence rate is small, units which have both large weights and the characteristic will have a greater impact on the variance and the design effects.

3.2.2 Provincial level

Provinces are also another important domain of interest. Knowing the sample allocation is not optimal at the provincial level, it is interesting to assess the variation of design effects at this level for the different weighting steps. Data was collected in the ten provinces and the three Territories, but since the Territories were excluded in the study, only the results for the 10 provinces are used in this study.

The following chart shows the median (P50) of design effects for four variables at the provincial level.

Chart 2: Median (P50) of design effects at the provincial level



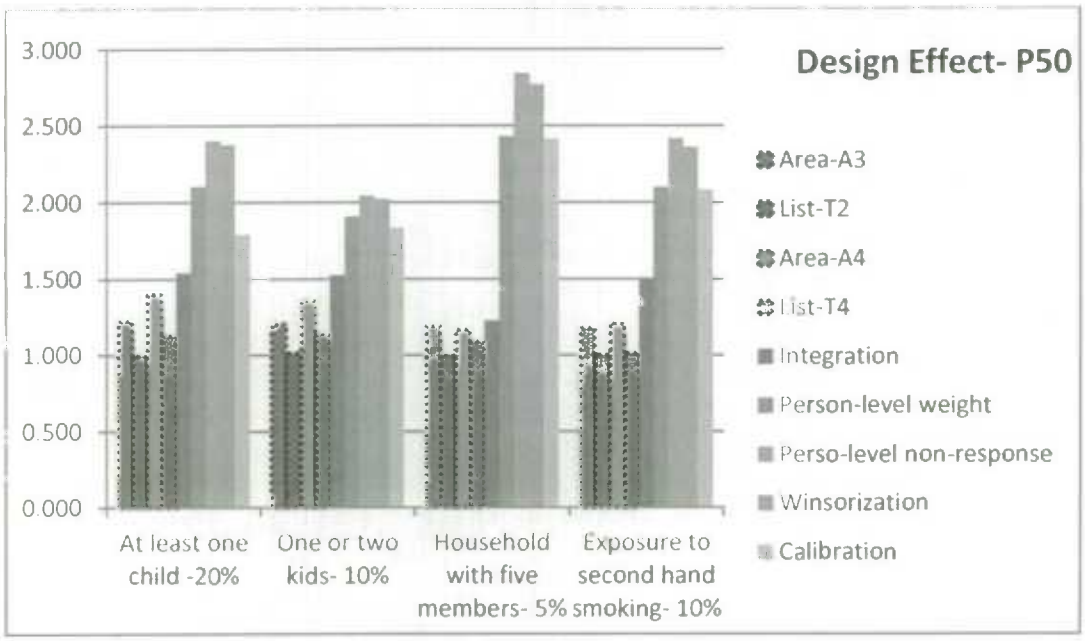
The trend of the design effects throughout the weighting process and by variable is similar to the results at the HR level. However, the design effects at the provincial level are higher than the HR estimates since the allocation is less optimal at the provincial level.

It is important to mention that even with higher design effects, the province-level estimates have higher quality than the ones at the HR level in terms of lower CVs due to larger sample sizes at the provincial level.

3.2.3 CCHS stratum level

The CCHS is often stratified at the HR level but in some instances a sub-HR level of geography is used. Tracking the changes of the design effect at this level shows the real impact of CCHS design. The CCHS considers 144 strata, however, after exclusion of the Territories, the RDD HRs and the strata with few units, only the results of 135 strata were considered in this study. The following chart shows the median (p50) of design effects for the four variables at the CCHS stratum level.

Chart 3: Median (P50) of design effects at the CCHS stratum level



The design effects of the variables at the CCHS stratum level are very similar to what we have at the HR level. In fact, minor differences between HR estimates and CCHS stratum estimates are due to the stratification at sub-HR level in some HRs.

4. CONCLUSION

This study has found which of the survey steps cause more variability in the estimates and cause the estimates to be less efficient compared to a simple random sample. Findings using 2010 CCHS data can be summarized as:

1. Using multiple frames, under current integration method with fixed α, increases the variability of design. (Integration Step-I1)

2. Using unequal person-selection probabilities has a negative effect on the design effect. (Person-level weight-I2)

3. The calibration using population counts significantly decreases the design effect.

4. The design effects at the person level will be higher for small prevalence, especially when people having the characteristic have large weight adjustments.

5. The design effects in the three considered domains (provincial level, HR level and CCHS stratum level) have similar trends throughout the weighting process.

Considering the above points, a redesigned CCHS should attempt to use only one frame which would ideally consist of a list of individuals instead of a list of dwellings. This would remove the need to integrate the frames, and the need to select one person per household, which are important factors that greatly affect the efficiency of the current design.

ACKNOWLEDGEMENTS

The authors would like to thank Jean-Francois Beaumont and Sharon Wirth for their review of this document.

Appendix A: Variable Definitions and Specifications

1. Households with at least one child ~ 20%

Child is defined as a person who is less than 12 years old and the variable is defined based on DHHDL12 which indicates the number of people living within a household whose age is less than 12 years old.

Numerator:

$$\text{At least one child} = \begin{cases} 1 & \text{if DHHDL12} >= 1 \\ 0 & \text{otherwise} \end{cases}$$

Denominator:

The denominator of this variable is defined as "All" which is equal to 1 for all inscope households

$$\%\text{ of households with at least one child} = \frac{\text{Households with at least one child}}{\text{All inscope households}}$$

2. Household with one or two kids~10%

Kids refer to person living in a household whose age is less than 6. This variable is derived by DHHDLE5 that indicates the number of people living within a household whose age is less than 6 years old.

Numerator:

$$\text{One or two kids} = \begin{cases} 1 \text{ if DHHDLE5} = 1 \text{ or } 2 \\ 0 \quad \text{otherwise} \end{cases}$$

DHHDLE5 is created by sorting the household roster dataset by SAMPLEID[1] and PERSONID[1] and by counting the number of PERSONIDs that have a DHH_AGE value less than 6 within each SAMPLEID

Denominator:

The denominator of this variable is defined as "All" which is equal to 1 for all inscope households.

$$\%\text{ of households with one or two kids} = \frac{\text{Households with one or two kids}}{\text{All inscope households}}$$

3. Household with five members~5%

This variable is defined based on DHHDHSZ that indicates the number of people living within a household (household size).

Numerator:

$$\text{Five members} = \begin{cases} 1 & \text{if DHHDHSZ=5} \\ 0 & \text{otherwise} \end{cases}$$

Denominator:

The denominator of this variable is defined as "All" which is equal to 1 for all inscope households.

$$\%\text{ of households with five members} = \frac{\text{households with five members}}{\text{All inscope households}}$$

4. Household with exposure to second-hand smoke~10%

This variable is defined based on the variable ETS_10 that indicates whether someone smokes inside the home including both household members and regular visitors. ETS_10 is applicable to respondents with DHHDHSZ > 1 or who answered (SMK_202 = (3, 7 or 8) or (SMK_01A = 8 and SMK_01B = 8))

Where:
DHHDHSZ is the household size.
SMK_202 is the type of smoker (At the present time, do you smoke cigarettes daily, occasionally or not at all?) SMK_202 = (3, 7 or 8) represent "Not at all", "Don't know" and "Refusal".
SMK_01A asks (In your lifetime, have you smoked a total of 100 or more cigarettes (about 4 packs)?) SMK_01A = 8 means Refusal
SMK_01B indicates whether someone has smoked a whole cigarette and SMK_01B = 8 means Refusal.

ETS_10 asks (Including both household members and regular visitors, does anyone smoke inside your home, every day or almost every day?)

1-Yes 2- No 6- Not Applicable 7-Don't know 8- Refusal 9- Not stated

Numerator:

$$\text{Exposure to second-hand smoke or SHSM} = \begin{cases} 1 & \text{if ETS\_10 = 1(yes)} \\ 0 & \text{otherwise} \end{cases}$$

Denominator:

Denominator of this variable is defined as "T_SHSM":

$$\text{T\_SHSM} = \begin{cases} 1 & \text{if ETS\_10 < 6 (Yes or No)} \\ 0 & \text{otherwise} \end{cases}$$

$$\%\text{ of households with exposure to second-hand smoke} = \frac{\text{Households with SHSM}}{\text{Households with T\_SHSM}}$$

Note that before the household nonresponse weighting step (before steps A4 and T3), households include households with reported AND imputed data in the calculation. At the nonresponse weighting step and after, households refer only to responding households in the calculation. This applies for each variable described above.

Appendix B: Some notes on the imputation of nonrespondent households

The nonrespondent households are imputed based on the prevalence of observed units at the CCHS stratum level. In other words, for variable v1 we counted the number of units with v1=1 and divide this number by total number of responding households in each CCHS stratum.

$$Pi=\frac{Number\ of\ Responding\ housholds\ with\ v1=1\ in\ CCHS\ stratum\ i}{Number\ of\ responding\ households\ in\ CCHS\ stratum\ i}$$

After computing Pi for all the CCHS strata, we generated a random number from uniform distribution U (0, 1) for each nonresponding household. If the generated random number was less than Pi then the household was imputed by v1=1 otherwise by v1=0.

Note: For the variable "Household with exposure to second-hand smoke", the variable imputed was ETS_10 by one of its potential values (1, 2, 6, 7, 8 or 9). Then, the derived variables SHSM and T_SHSM were computed based on the ETS_10 imputed value (See Appendix A for further information.)

Appendix C: Percentile of design effects at different domains of interest (HR, province and CCHS stratum) for each of the four variables and each weighting step

Table C1: Design Effects at the HR level- % of households with at least one child~ 20%

Weighting Step	Frame	Design Effect					
		Min	P 25	P 50	P 75	Max	Mean
Removal of out-of-scope units (A3 and T2)	**Area**	0.46	1.08	1.25	1.6	4.64	1.38
	Tel	0.87	0.99	1.03	1.07	1.33	1.04
Household Nonresponse (A4 and T3)	**Area**[1]	**0.41**	**1.2**	**1.43**	**1.71**	**3.72**	**1.52**
		0.41	1.2	1.43	1.73	3.72	1.52
	Tel	0.92	1.09	1.15	1.23	1.5	1.16
Integration I1	**Combined**	0.82	1.29	1.58	1.99	4.35	1.7
Person-Weight I2	**Combined**	1.03	1.78	2.2	2.73	4.93	2.32
Person-nonresponse I3	**Combined**	1.21	2.02	2.43	3.1	6.09	2.67
Winsorization- I4	**Combined**	1.21	1.96	2.4	3.05	6.06	2.6
Calibration- I5	**Combined**	0.81	1.45	1.76	2.28	3.66	1.87

Table C2: Design Effect at the HR level- % of households with one or two kids~ 10%

Weighting Step	Frame	Design Effect					
		Min	P 25	P 50	P 75	Max	Mean
Removal of out-of-scope units (A3 and T2)	**Area**	0.29	1.04	1.23	1.55	2.95	1.33
	Tel	0.86	0.96	1.02	1.07	1.37	1.02
Household Nonresponse (A4 and T3)	**Area**[2]	**0.43**	**1.07**	**1.40**	**1.66**	**3.07**	**1.43**
		0.43	1.07	1.40	1.67	3.07	1.44
	Tel	0.90	1.07	1.15	1.21	1.76	1.15
Integration I1	**Combined**	0.90	1.29	1.57	2.00	3.94	1.68
Person-Weight I2	**Combined**	1.03	1.61	1.95	2.50	4.55	2.11
Person-nonresponse I3	**Combined**	1.07	1.69	2.10	2.76	7.91	2.34
Winsorization- I4	**Combined**	1.07	1.66	2.05	2.68	7.91	2.27
Calibration- I5	**Combined**	0.94	1.47	1.81	2.20	4.16	1.90

[1&2] bold numbers indicate design effects before undercoverage adjustment. (section 3.1.2)

Table C3: Design Effect at the HR level- % of households with five members~5%

Weighting Step	Frame	Design Effect					
		Min	**P 25**	**P 50**	**P 75**	**Max**	**Mean**
Removal of out-of-scope units (A3 and T2)	**Area**	0.23	0.97	1.20	1.43	3.96	1.26
	Tel	0.85	0.95	1.00	1.04	1.51	1.01
Household Nonresponse (A4 and T3)	**Area**[3]	**0.69**	**1.00**	**1.17**	**1.47**	**3.30**	**1.29**
		0.55	1.00	1.17	1.47	3.30	1.29
	Tel	0.81	1.02	1.10	1.18	1.44	1.10
Integration I1	**Combined**	0.74	1.03	1.26	1.66	5.33	1.40
Person-Weight I2	**Combined**	1.09	2.00	2.39	3.21	6.56	2.68
Person-nonresponse I3	**Combined**	1.32	2.21	2.84	3.65	7.14	3.05
Winsorization- I4	**Combined**	1.32	2.17	2.78	3.46	6.05	2.95
Calibration- I5	**Combined**	1.13	1.92	2.38	3.05	4.73	2.57

Table C4: Design Effect at the HR level- % of households with exposure to second-hand smoke~10%

Weighting Step	Frame	Design Effect					
		Min	**P 25**	**P 50**	**P 75**	**Max**	**Mean**
Removal of out-of-scope units (A3 and T2)	**Area**	0.55	1.00	1.18	1.43	3.49	1.28
	Tel	0.84	0.97	1.02	1.06	1.20	1.02
Household Nonresponse (A4 and T3)	**Area**[4]	**0.48**	**1.07**	**1.27**	**1.60**	**3.13**	**1.38**
		0.48	1.07	1.27	1.60	3.13	1.39
	Tel	0.79	1.04	1.13	1.21	1.50	1.13
Integration I1	**Combined**	0.78	1.26	1.55	1.92	4.00	1.61
Person-Weight I2	**Combined**	0.80	1.72	2.16	2.85	9.85	2.65
Person-nonresponse I3	**Combined**	0.87	1.87	2.50	3.36	17.66	2.98
Winsorization- I4	**Combined**	0.89	1.81	2.38	3.27	9.61	2.82
Calibration- I5	**Combined**	1.18	1.76	2.26	2.69	5.97	2.40

[3&] bold numbers indicate design effects before undercoverage adjustment. (section 3.1.2)

Table C5: Design Effect at the Provincial level- % of households with at least one child~ 20%

Weighting Step	Frame	Design Effect					
		Min	**P 25**	**P 50**	**P 75**	**Max**	**Mean**
Removal of out-of-scope units (A3 and T2)	**Area**	1.20	1.35	1.58	2.32	2.38	1.72
	Tel	1.00	1.07	1.21	1.33	1.54	1.22
Household Nonresponse (A4 and T3)	**Area**[5]	**1.32**	**1.43**	**1.77**	**2.27**	**2.80**	**1.89**
		1.31	1.39	1.75	2.31	2.61	1.85
	Tel	1.11	1.15	1.25	1.55	1.68	1.33
Integration I1	**Combined**	1.17	1.51	2.04	2.66	3.15	2.10
Person-Weight I2	**Combined**	1.60	2.03	2.94	3.97	4.17	2.90
Person-nonresponse I3	**Combined**	2.09	2.62	3.26	4.63	5.28	3.46
Winsorization- I4	**Combined**	1.96	2.07	3.15	4.22	4.98	3.18
Calibration- I5	**Combined**	1.24	1.74	2.21	2.67	3.30	2.23

Table C6: Design Effect at the Provincial level- % of households with one or two kids~ 10%

Weighting Step	Frame	Design Effect					
		Min	**P 25**	**P 50**	**P 75**	**Max**	**Mean**
Removal of out-of-scope units (A3 and T2)	**Area**	0.86	1.24	1.63	1.81	2.79	1.61
	Tel	0.97	1.05	1.20	1.25	1.34	1.17
Household Nonresponse (A4 and T3)	**Area**[6]	**0.97**	**1.53**	**1.65**	**2.03**	**2.97**	**1.75**
		0.93	1.49	1.65	1.96	2.80	1.72
	Tel	1.03	1.22	1.34	1.40	1.54	1.31
Integration I1	**Combined**	1.12	1.55	2.09	2.68	3.02	2.07
Person-Weight I2	**Combined**	1.91	2.72	2.93	3.22	3.58	2.84
Person-nonresponse I3	**Combined**	1.92	2.08	2.92	3.41	4.30	2.86
Winsorization- I4	**Combined**	1.41	1.97	2.80	3.26	3.48	2.62
Calibration- I5	**Combined**	1.17	1.76	2.31	2.56	2.68	2.14

[5&6] bold numbers indicate design effects before undercoverage adjustment. (section 3.1.2)

Table C7: Design Effect at the Provincial level- % of households with five members~ 5%

Weighting Step	Frame	Design Effect					
		Min	**P 25**	**P 50**	**P 75**	**Max**	**Mean**
Removal of out-of-scope units (A3 and T2)	**Area**	0.65	1.35	1.62	1.66	1.81	1.50
	Tel	0.97	1.06	1.11	1.26	1.43	1.16
Household Nonresponse (A4 and T3)	**Area**[7]	**0.64**	**1.39**	**1.76**	**1.85**	**1.96**	**1.59**
		0.63	1.35	1.76	1.80	1.90	1.58
	Tel	1.07	1.15	1.26	1.41	1.55	1.28
Integration I1	**Combined**	1.01	1.40	1.77	2.07	2.33	1.73
Person-Weight I2	**Combined**	2.06	2.89	3.32	3.99	5.86	3.61
Person-nonresponse I3	**Combined**	2.23	3.33	4.22	4.67	6.49	4.06
Winsorization- I4	**Combined**	2.09	3.04	3.58	4.52	5.11	3.61
Calibration- I5	**Combined**	2.06	2.36	2.88	3.60	4.45	2.98

Table C8: Design Effect at the Provincial level- % of households with exposure to second-hand smoke~10%

Weighting Step	Frame	Design Effect					
		Min	**P 25**	**P 50**	**P 75**	**Max**	**Mean**
Removal of out-of-scope units (A3 and T2)	**Area**	0.82	1.35	1.41	1.52	2.19	1.44
	Tel	0.99	1.01	1.08	1.22	1.41	1.12
Household Nonresponse (A4 and T3)	**Area**[8]	**0.84**	**1.36**	**1.55**	**1.62**	**2.15**	**1.50**
		0.80	1.34	1.43	1.48	1.96	1.41
	Tel	0.97	1.02	1.09	1.22	1.39	1.13
Integration I1	**Combined**	1.26	1.53	1.73	2.04	2.41	1.76
Person-Weight I2	**Combined**	1.72	2.41	2.67	3.78	4.57	3.08
Person-nonresponse I3	**Combined**	2.09	2.62	3.26	4.63	5.28	3.46
Winsorization- I4	**Combined**	1.64	2.48	2.88	3.47	4.37	3.02
Calibration- I5	**Combined**	2.06	2.36	2.88	3.60	4.45	2.98

[7&8] bold numbers indicate design effects before undercoverage adjustment. (section 3.1.2)

Table C9: Design Effect at the CCHS stratum level - % of households with at least one child~ 20%

Weighting Step	Frame	Design Effect					
		Min	P 25	P 50	P 75	Max	Mean
Removal of out-of-scope units (A3 and T2)	**Area**	0.00	1.03	1.23	1.60	4.64	1.35
	Tel	0.85	0.96	1.00	1.04	1.15	1.00
Household Nonresponse (A4 and T3)	**Area**[9]	**0.03**	**1.15**	**1.40**	**1.72**	**5.02**	**1.50**
		0.03	1.15	1.40	1.72	5.02	1.50
	Tel	0.92	1.07	1.13	1.21	1.44	1.14
Integration I1	**Combined**	0.66	1.24	1.54	1.99	5.83	1.72
Person-Weight I2	**Combined**	0.95	1.74	2.10	2.71	6.92	2.34
Person-nonresponse I3	**Combined**	0.94	1.94	2.40	3.03	6.70	2.67
Winsorization- I4	**Combined**	0.94	1.91	2.38	2.98	6.68	2.60
Calibration- I5	**Combined**	0.85	1.48	1.79	2.43	6.46	2.01

Table C10: Design Effect at the CCHS stratum level - % of households with one or two kids~ 10%

Weighting Step	Frame	Design Effect					
		Min	P 25	P 50	P 75	Max	Mean
Removal of out-of-scope units (A3 and T2)	**Area**	0.29	0.99	1.20	1.45	5.91	1.30
	Tel	0.87	0.96	1.01	1.06	1.36	1.01
Household Nonresponse (A4 and T3)	**Area**[10]	**0.43**	**1.02**	**1.36**	**1.62**	**4.94**	**1.41**
		0.43	1.02	1.36	1.62	4.94	1.41
	Tel	0.86	1.07	1.14	1.20	1.43	1.13
Integration I1	**Combined**	0.75	1.23	1.52	1.90	7.27	1.68
Person-Weight I2	**Combined**	0.85	1.50	1.91	2.45	7.11	2.09
Person-nonresponse I3	**Combined**	0.88	1.60	2.04	2.74	7.88	2.31
Winsorization- I4	**Combined**	0.88	1.60	2.02	2.59	7.88	2.25
Calibration- I5	**Combined**	0.96	1.48	1.83	2.22	6.94	1.95

[9]& bold numbers indicate design effects before undercoverage adjustment. (section 3.1.2)

Table C11: Design Effect at the CCHS stratum level - % of households with five members~ 5%

Weighting Step	Frame	Design Effect					
		Min	**P 25**	**P 50**	**P 75**	**Max**	**Mean**
Removal of out-of-scope units (A3 and T2)	**Area**	0.23	0.97	1.19	1.50	3.96	1.27
	Tel	0.85	0.96	1.00	1.04	1.15	1.00
Household Nonresponse (A4 and T3)	**Area**[11]	**0.61**	**1.00**	**1.17**	**1.47**	**3.30**	**1.29**
		0.61	1.00	1.17	1.47	3.30	1.29
	Tel	0.83	1.02	1.09	1.18	1.39	1.10
Integration I1	**Combined**	0.65	1.01	1.23	1.65	4.89	1.38
Person-Weight I2	**Combined**	1.09	1.93	2.43	3.19	6.54	2.66
Person-nonresponse I3	**Combined**	1.36	2.15	2.84	3.53	7.11	3.01
Winsorization- I4	**Combined**	1.36	2.10	2.77	3.40	5.98	2.91
Calibration- I5	**Combined**	1.14	1.96	2.41	3.05	4.72	2.57

Table C12: Design Effect at the CCHS stratum level - % of households with exposure to second-hand smoke~10%

Weighting Step	Frame	Design Effect					
		Min	**P 25**	**P 50**	**P 75**	**Max**	**Mean**
Removal of out-of-scope units (A3 and T2)	**Area**	0.41	0.97	1.18	1.42	3.65	1.25
	Tel	0.87	0.97	1.00	1.06	1.20	1.01
Household Nonresponse (A4 and T3)	**Area**[12]	**0.42**	**0.95**	**1.21**	**1.47**	**3.70**	**1.28**
		0.42	0.95	1.21	1.47	3.70	1.28
	Tel	0.88	0.97	1.02	1.06	1.25	1.02
Integration I1	**Combined**	0.83	1.18	1.50	1.87	4.02	1.57
Person-Weight I2	**Combined**	0.86	1.70	2.10	2.68	9.81	2.55
Person-nonresponse I3	**Combined**	0.84	1.86	2.42	3.14	17.40	2.89
Winsorization- I4	**Combined**	0.83	1.83	2.36	3.11	9.55	2.75
Calibration- I5	**Combined**	0.76	1.74	2.08	2.72	6.04	2.38

[11&12] bold numbers indicate design effects before post-stratification. (section 3.1.2)

REFERENCES

Kish, Leslie (1965), *Survey Sampling*, John Wiley and Sons, New York.

Rao J. N. K., and Wu C. F. J. (1988), "Resampling Inference with Complex Survey Data", Journal of the American Statistical Association, 83, pp. 231-241.

Sautory O. (2003) CALMAR 2: A New Version of the CALMAR Calibration Adjustment Program. *Proceedings of Statistics Canada Symposium* (Statistics Canada, Catalogue no. 11-522-XCB), 2003.

Statistics Canada (2011), "CCHS - 2010 and 2009-2010 Microdata User Guide", *Statistics Canada*.

Statistics Canada (2011), 'Bootstrap Variance Estimation with the Canadian Community Health Survey', *Statistics Canada*.

Wilder Kate and Thomas Steven (2010), "Dual frame Weighting and Estimation Challenges for the Canadian Community Health Survey", *Proceedings of Statistics Canada, Symposium 2010*.