STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 79—94

NONRESPONSE BIAS IN THE SURVEY OF YOUTH UNDERSTANDIG OF SCIENCE AND TECHNOLOGY IN BOGOTÁ

Edgar Mauricio Bueno Castellanos[1]

ABSTRACT

The Colombian Observatory of Science and Technology -OCyT- developed, in 2009, a survey about understanding of Science and Technology in students of high school in Bogotá, Colombia. The sampling design was stratified according to the nature of school (public or private). Two sources of unit nonresponse were detected. The first one corresponds to schools that did not allowed to collect information. The second source corresponds to students who did not assist during the days when survey was applied. Estimates were obtained through two different approaches. Results obtained in both cases do not show visible differences when estimating ratios; even though, some great differences were observed when estimating totals. Results obtained using the second approach are believed to be more reliable because of the methodology used to handle item nonresponse.

Key words: Sampling design; nonresponse bias; calibration.

1. Introduction

In 2009, the Colombian Observatory of Science and Technology -OCyT- developed the *Survey of Youth Understanding of Science and Technology in Bogotá*, which inquires about topics related to understanding about scientist, engineers and benefits and risks of science and technology. Results and analysis of the survey are presented by Daza et. al (2011).

As expected, on the data collecting process, there were students who were not possible to contact (unit nonresponse) and others that did not fulfill some of the questions in the questionnaire (item nonresponse). As a consequence, arises the need to use methodologies that allows to obtain estimations taking into account the presence of nonresponse.

[1] Colombian Observatory of Science and Technology –OCyT-. Bogotá Colombia.
E-mail: embuenoc@ocyt.org.co.

Initially, item nonresponse was considered as a new category and the unit nonresponse was handled by conforming Response Homogeneity Groups (Särndal, Swensson and Wretman, 1992). After that, it was proposed to obtain estimations through other methodology: to impute missing values corresponding to item nonresponse and to use the calibration estimator for unit nonresponse.

The second section of this document describes the methodology used for design and development of the survey. The third section describes the causes of nonresponse in the survey and the two methodologies proposed to handle it. In order to compare these methodologies, a Monte Carlo simulation was carried out, its results are described in the fourth section. This simulation allowed to see the behavior of estimators under different cases. In the last section conclusions and suggestions are presented based on the experience achieved through the survey.

2. Methodology

The survey target population was conformed to students of the last two years of high school of all the schools in Bogotá, Colombia. The sampling frame used to identify the schools was the educational establishment register from the Secretaría de Educación de Bogotá (bureau of education), which includes, besides identification and contact variables, the nature of school (public or private) and information about the number of students registered in year 2008 in every grade. The register includes all the educational establishments in the city, therefore, it was necessary to eliminate the institutions that does not offer the grades defined for the study and those that offer them but have an approach on adult education. Finally, it was obtained a sampling frame with 1073 schools, 715 of these are private and reported 59984 students in 2008, the remaining 358 are public and reported 112830 students in the same year.

Once conformed the final frame, the sample was drawn. The design was a Stratified one-stage cluster sampling.

- The nature of school was used as stratification variable.
- In each stratum a sample of schools was drawn using a probability proportional to size *-pps-* design. The size variable used to assign probabilities to schools was the number of students reported for 2008 according to the frame, incremented in one unity.
- The questionnaire was applied to every student in the last two years in selected schools.

A sample of 31 private and 16 public schools was drawn (ordered sample). One public and two private schools were reselected in the sample, obtaining a set-sample of 29 private and 15 public schools, which have, respectively, 6231 and 7498 students in 2008. Throughout this document and unless otherwise is specified, *sample* will make reference to the ordered sample.

When the data collection stage ended, paper questionnaires were transcribed, conforming a data set that was validated and then estimations were carried out. In a first moment, it was planned to obtain estimations using the estimator proposed by Hansen and Hurwitz (1943), also known as with-replacement sampling estimator –pwr estimator-.

This estimator could not be used because it was not possible to obtain information from all individuals expected in the sample. For this reason it was necessary to identify other alternatives to obtain estimations in the presence of nonresponse. The next section describes the alternatives used for the survey.

3. Dealing with nonresponse

As usual when developing a survey, in the understanding survey both types of nonresponse arises: item nonresponse and unit nonresponse.

Two unit nonresponse sources were identified. The first one, due to directives that deny data collection: the survey was implemented in the 16 public schools, but only in 13 out of 31 private schools drawn in the sample; this case will be referred as *cluster nonresponse*. The second source corresponds to students who belong to schools in which access was allowed but did not assist during the days when survey was applied, this case will be referred as *element nonresponse.*

Given that all the questions in the survey are categorical, in a first moment estimations were obtained by considering item nonresponse as a new category for every variable. Unit nonresponse was handled by modifying expected sample sizes by those observed. This approach will be referred as *Approach 1*.

Later, it was decided to obtain new estimations by the use of methods allowing to control the nonresponse effects: the *nearest neighbor* methodology was used to impute values belonging to item nonresponse and the calibration estimator was used to handling unit nonresponse. This approach will be referred as *Approach 2* and is described in Section 3.2.

a. Approach 1

In this approach the nonresponse was handled according to:

Item nonresponse: Missing values due to item nonresponse was considered as a new category. By doing this, a rectangular data set is obtained, in which missing values are replaced by a code representing its absence. One advantage of the methodology is that allows to obtain a completely rectangular data set allowing to make cross tabulation of variables in survey; on the other hand, some disadvantages are the arising of meaningless cross-classified cells and that nothing is done in order to control the bias due to nonresponse.

Unit nonresponse: Element nonresponse was handled by assuming that, in every school, students who participated in the survey conform a simple random sample of students. The bias generated by this assumption is expected to be small given that the nonresponse rates within the schools that participate in the survey were low.

Cluster nonresponse was handled by assuming a *response homogeneity group* model with groups given by the nature of school. This means that is assumed that the response probability, θ_i, in each group of schools (public or private) is fixed and estimated by $\hat{\theta}_i = m_h^*/m_h$. In this case, *pwr* estimator takes the form

$$\hat{t}_y^* = \sum\nolimits_h \left(\frac{1}{m_h^*}\sum\nolimits_{r_h}\frac{\hat{t}_{yi}}{p_i}\right), \quad h = 1{,}2; \text{ with } \hat{t}_{yi} = \frac{N_i}{n_i}\sum\nolimits_{r_i} y_k \tag{1}$$

where

- $\hat{t}_{yi}$ is the estimation of the total of variable y in the ith school,
- N_i is the number of students in the ith school. This number was recorded for every school in the response set,
- n_i is the number of students who answered the questionnaire in the ith school,
- r_i is the response set of students belonging to the ith school,
- r_h is the response set of schools in the stratum h,
- m_h is the number of selected schools in stratum h,
- m_h^* is the number of schools in the response set in stratum h,
- y_k is the value of y for the kth,
- p_i is the selection probability of ith school.

At first glance, the estimator in (1) does not control the bias or the variance increments that may be generated as a consequence of the nonresponse. Even so, this estimator satisfies the desirable property of reproducing totals for the size variable used to obtain the selection probabilities of individuals:

For the case of element sampling from a population U, which counts with values of y for every element in the sample s of size m, let $t_x = \sum_U x_k$ the total of size variable x and x_k the value of x associated to the kth individual. Selection probability for kth individual is defined as $p_k = x_k/t_x$. When applying the pwr estimator to values of x in the sample s, we obtain

$$\hat{t}_x = \frac{1}{m}\sum\nolimits_s \frac{x_k}{p_k} = t_x$$

This result, obtained for element sampling, works also for the modification proposed for handling nonresponse, equation (1),

$$\hat{t}_{h,x}^* = \frac{1}{m_h^*}\sum\nolimits_{r_h}\frac{\hat{t}_{x_i}}{p_i} = t_{h,x}$$

This property indicates that, if y = αx, the estimator given in (1) will obtain perfect estimations for t_y for every sample, no matters the nonresponse, this property resembles the calibration estimator. It is clear that is impossible that the proportionality be satisfied in practice, even so, the property suggest that while there exists a high correlation between y and x, both, variance and bias due to nonresponse will be small. For the understanding survey it is required that totals in schools ($\mathrm{t_{y_i}}$) to be proportional to the number of students reported for 2008 ($\mathrm{t_{x_i}}$).

b. Approach 2

The nonresponse was handled according to:

Item nonresponse: Missing values due to item nonresponse were imputed using the *nearest neighbor* methodology: For every variable in the questionnaire, y_j, a set of G variables W, which is expected to be related to y_j, is identified and then sorted according to *explanatory power* expected with y_j, this is, the first variable in W will be the one that explain the most of y_j, the second will be the one following this rule, and so on. It is important to clarify that every variable in the study was qualitative and that the choice of the variables in W and its order were due to subjective criteria.

Individuals in the data set were divided in two groups according to the values for $\mathrm{y_j}$: the response set $\mathrm{r_j}$ and the nonresponse set $\mathrm{r_u - r_j}$, where $\mathrm{r_u = \bigcup_{j=1}^{q} r_j}$ is the set of individuals having information for at least one of the q variables in the questionnaire. The value $\mathrm{y_{kj}}$ (in $\mathrm{r_u - r_j}$) is imputed as follows:

- The matrix Z is created from W as: $z_{lg} = \begin{cases} 1, & \text{if } w_{kg} = w_{lg} \\ 0, & \text{if } w_{kg} \neq w_{lg} \end{cases}, g = 1,2,\cdots,G;\ l = 1,2,\cdots,n_{rj}$, where n_{rj} is the number of individuals in r_j.
- For every individual, l, in Z we calculate $D_k(l) = 2^{G-1}z_{l1} + 2^{G-2}z_{l2} + \cdots + 2^0 z_{lG}$.
- Individual that maximizes $D_k(l)$ is identified and its value y_{lj} is assigned to y_{kj}.
- When there are ties, y_{kj} is obtained as the mode of the y values associated to those individuals that maximizes $D_k(l)$.
- If there is not a unique mode, a random value of y is chosen from the set of modes.

It is clear that the distance metric $\mathrm{D_k(l)}$ is such that matching in $\mathrm{z_{l1}}$ dominates matching in the remaining variables $\mathrm{z_{l2}}$ to $\mathrm{z_{lG}}$; if $\mathrm{z_{l1}}$ does not match, $\mathrm{z_{l2}}$ dominates matching in the remaining $\mathrm{z_{l3}}$ to $\mathrm{z_{lG}}$; and so on. This situation was decided in order of reducing the burden of calculations that would imply the assignation of different weights to every variable in W for every variable in the questionnaire.

Unit nonresponse: Element nonresponse was handled in the same fashion that in *Approach 1:* it was assumed that, in every school, students who participated in the survey conform a simple random sample from the total of students.

Särndal and Lundström (2005) proposed the calibration estimator for the Horvitz and Thompson estimator (1952) at the level of individuals. For cluster nonresponse a variation of this estimator was used. Due to the absence of auxiliary information at the level of students, calibration was carried out at the level of schools using one quantitative and two categorical variables for classification.

Quantitative variable, $\mathrm{t}_{\mathrm{x}_\mathrm{i}}$, is the total of students in the ith school during 2008. The first classification variable, γ, is the same variable used for stratification, the nature of school (public or private), $\gamma_\mathrm{i} = (\gamma_{1\mathrm{i}}, \gamma_{2\mathrm{i}})'$, where

$$\gamma_{1i} = \begin{cases} 1, & \text{if school } i \text{ is private} \\ 0, & \text{if school } i \text{ is public} \end{cases} \text{ and } \gamma_{2i} = \begin{cases} 1, & \text{if school } i \text{ is public} \\ 0, & \text{if school } i \text{ is private} \end{cases}$$

The second classification variable, δ, is an indicator of the size of school, defined as $\delta_\mathrm{i} = (\delta_{1\mathrm{i}}, \delta_{2\mathrm{i}}, \delta_{3\mathrm{i}})'$, where

$$\delta_{1i} = \begin{cases} 1, \text{ if } t_{x_i} \leq 100 \\ 0, \text{ otherwise} \end{cases}, \delta_{2i} = \begin{cases} 1, \text{ if } 100 < t_{x_i} \leq 400 \\ 0, \text{ otherwise} \end{cases} \text{ and } \delta_{3i} = \begin{cases} 1, \text{ if } t_{x_i} > 400 \\ 0, \text{ otherwise} \end{cases}$$

The auxiliary vector associated to the ith school, t_{x_i}, is conformed as

$$\boldsymbol{t}_{x_i} = \left(\gamma_{1i}t_{x_i}, \gamma_{2i}t_{x_i}, \delta_{1i}t_{x_i}, \delta_{2i}t_{x_i}\right)'$$

and the input vector required is the total of students in every group in 2008:

$$\boldsymbol{X} = \left(\sum\nolimits_U \gamma_{1i}t_{xi}, \sum\nolimits_U \gamma_{2i}t_{xi}, \sum\nolimits_U \delta_{1i}t_{xi}, \sum\nolimits_U \delta_{2i}t_{xi}\right)'$$

δ_3 is not included in order to avoid singularities in the matrix to be inverted to obtain the calibrated selection probabilities.

Once defined the auxiliary vector and the input vector, the calibrated selection probabilities, w_i, are calculated as

$$w_i = p_i/v_i \text{ with } v_i = 1 + \boldsymbol{\lambda}_r' \mathbf{t}_{x_i} \text{ and}$$

$$\lambda_r' = \left(X - \sum\nolimits_h \left(\frac{1}{m_h}\sum\nolimits_{r_h} \frac{t_{x_i}}{p_i}\right)\right)\left(\sum\nolimits_h \left(\frac{1}{m_h}\sum\nolimits_{r_h} \frac{t_{x_i}t_{x_i}'}{p_i}\right)\right)^{-1}$$

and then, the total of y is estimated as

$$\hat{t}_y^c = \sum_h \left(\frac{1}{m_h}\sum_{r_h}\frac{\hat{t}_{y_i}}{w_i}\right),\ h = 1{,}2;\ \text{with } \hat{t}_{y_i} = \frac{N_i}{n_i}\sum_{r_i} y_k \qquad (2)$$

A comparison between the estimators $\hat{t}_y^*$ and $\hat{t}_y^c$ is presented in the next section.

4. A Monte Carlo simulation study

In order to compare the bias and variance of $\hat{t}_y^*$ and $\hat{t}_y^c$, defined in equations (1) and (2), respectively, a Monte Carlo simulation study was carried out. This process took into account only cluster nonresponse; element nonresponse and item nonresponse were ignored.

A population of $N = 148245$ individuals in 1073 schools was created. The number of schools was fixed to match the number of schools in the sampling frame, while the number of individuals was fixed to match the estimated number of students according to *Approach 1*.

Three auxiliary variables at the level of schools (x_1, x_2 and x_3), one *exogenous* variable (z) and three study variables at the level of students (y_1, y_2 and y_3) were generated as follows:

x_{i1}: Number of students in the ith school according to the sampling frame,

x_{2i}: The nature of ith school (public or private), $x_2 = \begin{cases} 1, & \text{if school is private} \\ 0, & \text{if school is public} \end{cases}.$

This variable is used also for conforming strata.

x_{3i}: The size of ith school, $x_3 = \begin{cases} 1, \text{if } x_1 \leq 100 \\ 2, \text{if } 100 < x_1 \leq 400. \\ 3, \text{if } x_1 > 400 \end{cases}$

z: A dichotomous exogenous variable related to the nature of school. By exogenous variable I mean a variable that is completely unknown in the survey: it is not an auxiliary variable known beforehand, and also is not measured in the questionnaire as a study variable:

$$P(z = 1|x_2 = 1) = 0.7 \qquad \text{and} \quad P(z = 1|x_2 = 0) = 0.4$$

y_1: A dichotomous variable that takes value 1 with different probabilities according to the nature of school:

$$P(y_1 = 1|x_2 = 1) = 0.8 \text{ and } \quad P(y_1 = 1|x_2 = 0) = 0.5$$

y_2: A dichotomous variable that takes value 1 depending on strata (x_2), and the value of z:

$P(y_2 = 1|x_2 = 1, z = 1) = 0.9,\quad P(y_2 = 1|x_2 = 1, z = 0) = 0.8,\quad P(y_2 =$
1|x2=0,z=1=0.5
and $\quad P(y_2 = 1|x_2 = 0, z = 0) = 0.2$

y_3: A dichotomous variable that takes value 1 depending only on the value of z:

$P(y_3 = 1|z = 1) = 0.9$ and $P(y_3 = 1|z = 0) = 0.45$

It is clarified that the auxiliary variables x_1, x_2 and x_3 are present at the level of schools, while the study variables y_1, y_2 and y_3 are present at the level of students and they are related to the auxiliary variables through its totals within schools.

The idea behind the setup for the study variables and the response distribution (step 4 of the simulation process) will be described below.

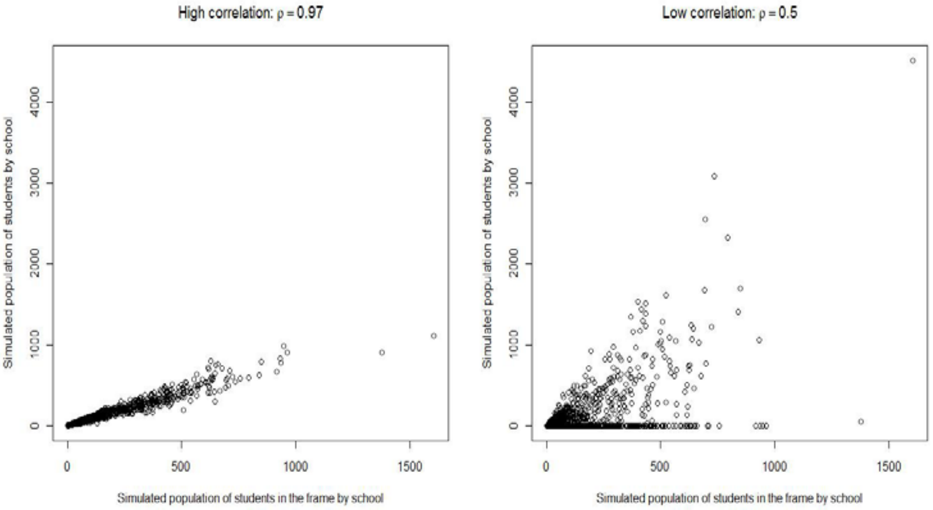

Figure 1. Simulated individuals by school: x_{1i} vs. N_i

The number of individuals in the ith school, N_i, was defined in order that the correlation coefficient between the size in the frame, x_{1i}, and N_i was (approximately) equal to ρ_0:

$$N_i = \hat{B}_0 x_{1i} + e_i\ ,$$

where $e_i = \varepsilon_i x_{1i}$ and ε_i is an observation of a random variable $N(0, a)$, with $a > 0$ chosen properly and $\hat{B}_0 = 0.32$ is the slope of the regression line of x_1 on N_i.

Two *correlation levels* between N_i and x_1 were generated: high ($\rho(N_i, x_{1i}) \approx 0.97$) and low ($\rho(N_i, x_{1i}) \approx 0.50$). The left panel of Figure 1 shows the scatter plot between the number of students for school according to the sampling frame and the number of students observed for the case $\rho(N_i, x_{1i}) \approx 0.97$. The right panel shows the case $\rho(N_i, x_{1i}) \approx 0.50$. It is clarified that the minimum value for the simulated populations is equal to 2.

With each of these populations the following process was carried out:

1. A stratified (with replacement) *pps* of 31 private and 16 public schools was drawn. The selection probability for the ith school was defined as $p_i = x_{1i}/\sum_{U_h} x_{1i}$. All students in selected schools were selected.
2. The totals by stratum of y_1, y_2, y_3 and the population size, N, from the full sample using the pwr estimator were estimated: $\hat{t}_{yh}^{(1)} = \frac{1}{m_h}\sum_{s_h} \frac{t_{y_i}}{p_i}$.
3. The totals by stratum of y_1, y_2, y_3 and the population size, N, using the calibration estimator including x_2 and x_3 as classification variables and x_1 as quantitative were estimated: $\hat{t}_{yh}^{(2)} = \frac{1}{m_h}\sum_{s_h} \frac{t_{y_i}}{w_i}$.
4. A school *response distribution* was generated by a fixed response probability θ_i depending on the strata and the school total of z. θ_i was defined in order that the response probability in Stratum 1 and Stratum 2, was 0.42 and 0.94, respectively.
5. Once defined the response set, totals for the four already mentioned variables were estimated using the estimator (1): $\hat{t}_{yh}^{(3)} = \frac{1}{m_h^*}\sum_{r_h} \frac{t_{y_i}}{p_i}$.
6. Totals of the four variables were estimated using the calibration estimator: $\hat{t}_{yh}^{(4)} = \frac{1}{m_h}\sum_{r_h} \frac{t_{y_i}}{w_i}$.
7. A stratified simple random sample -*srs*- of the full population of schools was drawn. This procedure was carried out in order to compare a design that includes auxiliary information (*pps*) against one that does not include it (*srs*). The number of schools, (m_h), was chosen with the goal that the number of individuals expected under the *srs* sample was (approximately) equal to the number of individuals expected under the *pps* sample.
8. Totals of y_1, y_2, y_3 and the population size, N, were estimated by using the Horvitz-Thompson estimator (also known as π-estimator) (1952): $\hat{t}_{yh}^{(5)} = \frac{M_h}{m_h}\sum_{s_h} t_{yi}$, with M_h the number of schools in stratum h.
9. In addition, with every estimator, the ratio $R_1 = t_{y1}/N$ was estimated. Also $R_2 = t_{y2}/N$ and $R_3 = t_{y3}/N$ were estimated. The results obtained are similar to those obtained for R_1.

The procedure described in numerals 1 to 9 is repeated $I = 10000$ times. Every time the estimations obtained through the five estimators are recorded. The (simulated) expectation of each estimator is obtained as

$$E_{SIM}\left[\hat{t}_y^{(j)}\right] = \frac{1}{I}\sum_{i=1}^{I} \hat{t}_{y_i}^{(j)}$$

and the (simulated) variance is obtained as

$$V_{SIM}\left[\hat{t}_y^{(j)}\right] = S_{\hat{t}_y^{(j)}}^2 = \frac{1}{I-1}\sum_{i=1}^{I}\left(\hat{t}_{yi}^{(j)} - E_{SIM}\left[\hat{t}_y^{(j)}\right]\right)^2.$$

Table 1 shows the parameters to estimate: totals of variables y_1, y_2 and y_3, total of individuals in the population and the ratio R_1.

Table 1. Population totals and ratios

	Total	Private schools	Official schools
N	148245	54417	93828
t_{y1}	90524	43664	46860
t_{y2}	77488	47452	30036
t_{y3}	100553	41370	59183
R_1	0,61	0,80	0,50

Tables 2 and 3 shows (simulated) relative bias and (simulated) coefficient of variation for cases $\rho(N_i, x_{1i}) \approx 0.97$ and 0.50, respectively. The (simulated) relative bias, RB_{SIM}, of the jth estimator for total t_y is calculated as

$$RB_{SIM}\left[\hat{t}_y^{(j)}\right] = \frac{E_{SIM}\left[\hat{t}_y^{(j)}\right] - t_y}{t_y}$$

and the (simulated) coefficient of variation, CV_{SIM}, of the jth estimator for total t_y is calculated as

$$CV_{SIM}\left[\hat{t}_y^{(j)}\right] = \frac{\sqrt{V_{SIM}\left[\hat{t}_y^{(j)}\right]}}{E_{SIM}\left[\hat{t}_y^{(j)}\right]}$$

A few words on the response distribution, the variables y_1, y_2 and y_3, and the auxiliary vector $\boldsymbol{x}_k$: According to Särndal and Lundström (2005) there is a triple $(\theta_k, y_k, \boldsymbol{x}_k)$ associated to every individual in the population. It is clear that, by construction, θ_k depends partially on the known $\boldsymbol{x}_k$ and partially on the unknown z. Given that y_{1k} depends only on x_{2k} not on z, in this case the nonresponse is completely explained by the auxiliary vector $\boldsymbol{x}_k$, so this case can be considered as Missing at Random -MAR-. y_{2k} depends on both x_{2k} and z, so the nonresponse is partially explained by x_{2k}. Finally, y_{3k} depends only on the unknown variable z, so the auxiliary vector is unable to explain the nonresponse distribution, at least directly.

Table 2 shows the results for the case in which the correlation between the number of individuals by school in the frame and the number of individuals observed by school is 0.97. About the bias, Table 2 suggests the following results:

- It is known that $\hat{t}_1$ and $\hat{t}_5$ are unbiased in total estimation and $\hat{t}_2$ is *asymptotically unbiased*. The simulation allows to see these facts. The bias of $\hat{t}_2$ and $\hat{t}_3$ is also small.
- In the stratum of high nonresponse (stratum 1) the bias for the calibration estimator under nonresponse ($\hat{t}_4$) although small, is notably greater than the bias for the *pwr* estimator under nonresponse ($\hat{t}_3$). Meanwhile, in the stratum 2, there is a reverse situation: bias of $\hat{t}_4$ is smaller than the bias of $\hat{t}_3$.
- The bias of the five estimators for the ratio R_1 are small.

Table 2. Simulated relative bias and simulated coefficient of variation (as a percentage) of five estimators for the case $\rho(N_i, x_{1i}) \approx 0.97$.

		Relative bias					Coefficient of variation				
Strata	Parameter	$t^{(1)}$	$t^{(2)}$	$t^{(3)}$	$t^{(4)}$	$t^{(5)}$	$t^{(1)}$	$t^{(2)}$	$t^{(3)}$	$t^{(4)}$	$t^{(5)}$
Stratum 1	N	0,02	0,00	-0,03	-0,21	-0,14	3,43	3,58	5,53	6,19	12,74
	t_{y1}	0,02	-0,01	-0,05	-0,21	-0,13	3,62	3,78	5,86	6,42	12,81
	t_{y2}	0,04	0,01	0,00	-0,14	-0,14	3,53	3,68	5,67	6,31	12,82
	t_{y3}	0,01	0,00	-0,07	-0,27	-0,13	3,67	3,82	5,90	6,48	12,71
	R_1	0,00	-0,01	-0,02	0,00	0,00	1,01	1,05	1,64	1,61	0,67
Stratum 2	N	-0,05	-0,10	-0,06	0,01	-0,13	5,18	5,27	5,37	5,50	14,67
	t_{y1}	-0,04	-0,10	-0,05	0,02	-0,12	5,57	5,66	5,72	5,84	14,68
	t_{y2}	-0,05	-0,13	-0,05	0,04	-0,15	5,80	5,90	6,00	6,18	14,82
	t_{y3}	-0,05	-0,10	-0,06	0,02	-0,15	5,46	5,55	5,62	5,76	14,72
	R_1	0,01	0,00	0,01	0,01	0,02	1,72	1,74	1,76	1,82	1,18

Some comments on the coefficients of variation in Table 2:

- The variance of the estimators for totals is highly reduced when including auxiliary information: CV of $\hat{t}_5$ are clearly greater than those of the other four estimators.
- The CV of the calibration estimator are slightly greater than those for the *pwr* estimator: the CV of $\hat{t}_2$ is slightly greater than the CV of $\hat{t}_1$, for the case of full response; and, the CV of $\hat{t}_4$ is slightly greater than the CV of $\hat{t}_3$, for the case of nonresponse.
- The CV of the estimators that works in the presence of nonresponse in the stratum 1 are clearly higher than those of the estimators that works under full response; on the other hand, in stratum 2 this difference is small. This is a consequence of the response probabilities in each stratum.
- When estimating the ratio R_1, the results differs from those for totals: in this case the smallest CV corresponds to the strategy (*srs*, π-estimator), a strategy that does not includes auxiliary information in the design or the estimation stage. This is due to the fact that when estimating totals, the size variable used in $\hat{t}_1$ to $\hat{t}_4$ is more or less related to the totals within schools, so reducing the variance; whereas, when estimating a ratio, the size variable does not explain the variation in the variable of interest, and can even cause a loss of efficiency with regard to a strategy that does not include auxiliary information.

Table 3. Simulated relative bias and simulated coefficient of variation (as a percentage) of five estimators for the case $\rho(N_i, x_{1i}) \approx 0.50$.

		Relative bias					Coefficient of variation				
Strata	Parameter	$t^{(1)}$	$t^{(2)}$	$t^{(3)}$	$t^{(4)}$	$t^{(5)}$	$t^{(1)}$	$t^{(2)}$	$t^{(3)}$	$t^{(4)}$	$t^{(5)}$
Stratum 1	N	-0,12	0,49	-17,53	-20,92	0,18	23,73	24,53	44,04	47,79	21,52
	t_{y1}	-0,12	0,48	-17,51	-20,91	0,18	23,78	24,58	44,12	47,87	21,57
	t_{y2}	-0,12	0,48	-17,55	-20,90	0,18	23,70	24,50	44,01	47,76	21,54
	t_{y3}	-0,13	0,48	-17,63	-21,02	0,18	23,72	24,52	44,07	47,78	21,46
	R_1	-0,01	-0,01	0,02	0,04	0,00	1,05	1,08	2,17	2,16	0,65
Stratum 2	N	0,34	0,24	1,34	1,43	0,30	31,23	32,28	32,11	33,97	32,31
	t_{y1}	0,35	0,24	1,35	1,44	0,30	31,11	32,15	31,99	33,83	32,21
	t_{y2}	0,33	0,21	1,33	1,44	0,27	31,21	32,25	32,10	34,00	32,46
	t_{y3}	0,34	0,25	1,35	1,45	0,30	31,27	32,32	32,15	34,04	32,40
	R_1	0,06	0,06	0,07	0,07	0,03	1,85	1,88	1,90	1,97	1,16

Table 3 show the results for the case when correlation between the number of individuals by school in the frame and the number of individuals observed is 0.50. In reference to the bias, it is observed that, in this case there is a strong bias in the two estimators that works under nonresponse ($\hat{t}_3$ and $\hat{t}_4$) in the stratum of high nonresponse (stratum 1). This is due to the fact that neither the design nor the auxiliary variables were (enough) correlated to the study variables, so they were unable to control the bias generated by the nonresponse. Even so, the bias for R_1 is still small.

Some comments on the variances in Table 3:

- The increment in the variance of all the estimators for totals when comparing with those in Table 2 is clear.
- The variance of the two estimators that works with auxiliary information and counts with a full sample, $\hat{t}_1$ and $\hat{t}_2$, is similar. This variance is, indeed, similar with that obtained for the estimator $\hat{t}_5$. This result suggest that in this case, the gain obtained by the size variable x_1 (at the design stage) and by the auxiliary vector (at the estimation stage) is *negligible* as a consequence of the low correlation between these and the study variables.
- The effect of nonresponse in the variance is visibly greater in the first stratum: compare $\hat{t}_3$ and $\hat{t}_4$ with $\hat{t}_1$ and $\hat{t}_2$, respectively. This result is a consequence of the low response probability in stratum 1 and the low correlation between the auxiliary variables and the response distribution.
- The variance of the estimators when estimating a ratio does not show an increment when comparing with the results in Table 2; once more, $\hat{t}_5$ is the estimator with the smaller variance.
- It is interesting that although y_1, y_2 and y_3 were generated under different conditions, the results are not affected by this fact. The explanation is that although z is not included directly in the survey, it is explained indirectly by the size of the schools.

In my opinion, the most interesting result that is obtained from the simulation study already described is that, although the nonresponse have visible effects in bias and variance of the estimators when estimating totals (effects that becomes even bigger when auxiliary information is not highly correlated with the survey variables), this weakness does not seem to be *inherited* when estimating a ratio: estimations are still reliable, no matter the presence or absence of *powerful* auxiliary information or the patterns imposed on the nonresponse distribution.

This is important for the understanding survey given that ratios (proportions) are the most important parameters to be estimated in it.

Two aspects were taken into account in order to make a choice on one of the approaches: handling of unit nonresponse (in terms of the bias and the variance in the simulation study) and the handling of item nonresponse. Finally, it was decided to choose the *Approach 2.* The reasons to make this choice are:

- With regard to the bias, both estimators have a similar behavior: when estimating totals the bias is small if there is a high correlation between the expected and observed number of students; on the other hand, the bias is equally great for both estimators when the correlation is low. When estimating a ratio (proportion) the bias is negligible for both estimators.
- With regard to the variance, again both estimators have a similar behavior: a small CV when there is a high correlation between the expected and observed number of students; a greater CV when this correlation is low and an even greater CV when there is a low nonresponse.
- With regard to the handling of item nonresponse, it is strongly believed that the methodology used in Approach 2 overcomes to that used in Approach 1, where little is done, while in Approach 2 the relation between study variables is used to impute the missing values.
- Handling of item nonresponse in *Approach 1* creates a new category for every variable, this category does not correspond to the original questionnaire, it is a consequence of an unlucky –although common- event: partially incomplete information on the responses of an individual. This new category is not a problem in *Approach 2*, in which final tables keeps the structure expected at the moment of the questionnaire design. This fact facilitates the results interpretation.

5. Conclusions

- Although it is clear that nonresponse is an undesirable, but almost inevitable event in any survey, in the understanding survey developed by the OCyT there was the fortune of identifying a variable associated with its occurrence: the nature of school. Given that this variable was considered since the design stage as a stratification variable, the effect that nonresponse could have on bias and variance was reduced.
- Estimations for totals yields clear differences between both approaches, moreover, *Approach 1* yields lower estimates than *Approach 2*. Even so, these

differences are reduced when estimating proportions. This result is consistent with the simulations carried out, where proportions were less sensitive to the estimator. Furthermore, estimations for ratios happened to be *insensitive* to design, estimator or response distribution.

- Although there was not auxiliary information available at the level of individuals, available variables at the level of school (nature of school, number of student in the last year) allowed to build estimators that reduced the effect of nonresponse on the final estimations.
- The *pwr* estimator for a probability proportional to size *-pps-* design resembles the calibration estimator in the sense that it *reproduces exactly* the total of the size variable. A consequence of this property is that, when the selection probabilities are highly correlated to the study variables, a reduction in bias and variance generated by nonresponse is obtained.
- The results from the simulation study shows that both estimators have similar behaviors and that achieves satisfactorily the goal of controlling bias and variance generated by nonresponse when there is a high correlation between the expected and the observed number of students in schools.
- Simulations shown in section 4 allow to see the behavior of both proposed estimators in a set of cases. These cases were proposed in the context of the understanding survey and they were useful to make decisions on the estimators. Even so, it is important to recall that results must not be generalized, since they depends on the simulated population, considered designs, auxiliary variables included, response distribution, and so on.

REFERENCES

DAZA, S. ED, (2011). Entre datos y relatos: percepciones de jóvenes escolarizados sobre la Ciencia y la Tecnología. Observatorio Colombiano de Ciencia y Tecnología, Bogotá.

HANSEN, M.H., and HURWITZ W.N. (1943). On the theory of sampling from finite populations. Annals of Mathematical Statistics 41, 517-529.

HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663-685.

R Development Core Team, (2011). A Language and environment for Statistical Computing, http_//www.r-project.org.

SÄRNDAL, C.E. and SWENSSON, B. and WRETMAN, J., (1992). Model Assisted Survey Sampling. Springer.

SÄRNDAL, C.E. and LUNDSTRÖM, S., (2005). Estimation in Surveys with Nonresponse. Wiley.