Working Paper Series, N.11, June 2005

A simple two-stage data collection procedure to handle unit non-response in social surveys

Anna Giraldo
Department of Statistical Sciences
University of Padua
Italy

Gianpiero Dalla Zuanna
Department of Statistical Sciences
University of Padua
Italy

Abstract: The aim of this paper is to investigate the possibility of collecting information on the reproductive behaviour of mothers by means of a two-stages data collection procedure. At first stage a simple self-filled questionnaire is administered to a sample of students attending compulsory school; thus, each pupil is asked to give his mother a more complex questionnaire, bringing it back to school. Linking the two questionnaires, the non-response selection problems are solved using the same statistical techniques used to control attrition in panel surveys. Even if self-filled questionnaires can not be too complex, in some cases this data collection procedure can be recommended as (1) the proportion of unit non-responses in standard surveys can be very high, and the unit non-response selection is not easy to control; (2) it is a very low cost data collection procedure. After a general description of this data collection technique, an application is performed using data of the Urban Fertility Survey (UFS) conducted during 2001-02 in four Italian towns. As respondents are clearly selected (e.g., by citizenship) if data analysis is performed without correcting for non-response, results may be biased. In particular, we compare some regression models explaining fertility using data collected by the two questionnaires.

Keywords: unit non-response, fertility

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padua
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Anna Giraldo
tel: +39 049 827 4168
anna.giraldo@unipd.it
http://www.stat.unipd.it/agiraldo

A simple two-stages data collection procedure to handle unit non-response in social surveys

Anna Giraldo
Department of Statistical Sciences
University of Padua
Italy

Gianpiero Dalla Zuanna
Department of Statistical Sciences
University of Padua
Italy

Abstract: The aim of this paper is to investigate the possibility of collecting information on the reproductive behaviour of mothers by means of a two-stages data collection procedure. At first stage a simple self-filled questionnaire is administered to a sample of students attending compulsory school; thus, each pupil is asked to give his mother a more complex questionnaire, bringing it back to school. Linking the two questionnaires, the non-response selection problems are solved using the same statistical techniques used to control attrition in panel surveys. Even if self-filled questionnaires can not be too complex, in some cases this data collection procedure can be recommended as (1) the proportion of unit non-responses in standard surveys can be very high, and the unit non-response selection is not easy to control; (2) it is a very low cost data collection procedure. After a general description of this data collection technique, an application is performed using data of the Urban Fertility Survey (UFS) conducted during 2001-02 in four Italian towns. As respondents are clearly selected (e.g., by citizenship) if data analysis is performed without correcting for non-response, results may be biased. In particular, we compare some regression models explaining fertility using data collected by the two questionnaires.

Keywords: unit non-response, fertility.

JEL-Code: C25, C81, J13.

1. Introduction

One of the most intriguing practical and theoretical problem in social surveys is the high rate of unit non-responses, i.e., sampled people of the main list being impossible to be contacted or refusing the interview. Let us consider UN-ECE Family and Fertility Survey conducted in several developed countries during the 90s and few important surveys conducted in Italy during the last years.

In Family and Fertility Surveys, the non-response rate ranges from 5% (women in Portugal) to 84% (women in Greece). The median level is about 25% for men and 20% for women (table 1).

In Italy, the Survey of Household Income and Wealth conducted by Bank of Italy (Banca d'Italia, 2000, 2002) – currently and intensively used by economical and social researchers – is characterized by a very high non-response rates: 56% in 1998 and 62% in 2000. Non-responses are lower for surveys of Istat (the Italian National Statistical Institute): 26% in the 2001 Family Expenditure Survey (Istat, 2002); 16% in the 2002 Multipurpose Survey[1], but also for these cases, the unit non-response rate is not negligible and might affect the results.

Unit non-response can affect precision of the estimate due to a reduced sample size (if non-respondents are not replaced), but it would not be a major problem if the characteristics of respondents and non-respondents were the same. Unluckily, it is very difficult to asses if it does, as information on non-respondents most times are poor or missing. However, if some additional data for both respondents and non respondents are available, it can be possible to compare the two sub-groups. If data on non-respondents are rich enough, is also possible to correct the collected data, in

[1] Personal communication to the authors, unpublished data.

order to represent – from the statistical viewpoint – the whole population, rather than the only respondent sub-population (Little and Rubin, 2002).

Table 1: Non-response rates in UN-ECE Family and Fertility Survey.

%	*Men*	*Women*
0-9	Poland, Portugal	Poland, Portugal
10-19	Estonia, Finland, France, Slovenia	Estonia, Finland, France, Hungary, Norway, Spain, Slovenia
20-29	Hungary, Latvia, Norway, Spain, Sweden, Germany	Belgium, Latvia, Sweden
30-39	Belgium, Lithuania	Lithuania
40-49	Italy	Italy, New Zeland*
50+	Greece, Netherlands, Switzerland	Greece, Netherlands, Switzerland

* Data collected only for women

Source: Festy and Prioux (2002), p. 21. For Italy see also De Sandre et. al. (1997).

The main aim of this paper is to illustrate a simple and cheap data-collection strategy and data-analysis procedure in order to control the unit non-response problem in surveys conducted in countries where education is universally widespread among young people. The principal role of this data-collection strategy is played by the couple parent-pupil, and data are gathered in both primary school and family. In order to implement this procedure, we need only that (1) a large part of boys and girls of the eligible age attend school regularly, and (2) pupils are able to report in the questionnaire fill-in at school some elementary but correct information on their parents or relatives (i.e., the person who should fill-in the questionnaire at home).

1.1. The Urban Fertility Survey

We briefly describe the research project that was the occasion for developing and testing these procedures of data-collection and unit non-response analysis. The *Urban Fertility Survey* (UFS) was conducted in the winter 2001-02 by a pool of demographers, economists, and statisticians in four Italian towns: Padova and Udine (North), Florence (Centre), Messina (South)[2]. UFS was composed by two surveys that shared a large sets of questions, but were conducted with different techniques of data-collection (for details see Dalla Zuanna, Salvini, 2001, 2003). The first one was a CATI survey, targeted to women aged 40-44 without children, sampled by Municipality Registers of the four towns. The target of the second survey – based on self-filled questionnaires – were mothers of pupils attending the last year of primary school (aged about 13), i.e., women aged about 42 with at least one child. In this paper, only the unit non-response problems of the second survey are taken into account.

1.2. Data collection: the couple pupil-mother

In Udine, Padua, Florence and Messina we extracted a random sample of junior high schools (the second level of compulsory school in Italy). All the deans accepted that their school participated to UFS. For each school, interviewers entered in all classes of last year (where rather all boys and girls were about 13 years old). A brief self-administrated questionnaire (5-8 minutes long, the child-questionnaire, **CQ**) was submitted to all pupils. They filled it in immediately, and the interviewer exited from the school with all the **CQ**. A longer questionnaire (30-40 minutes, the mother-questionnaire, **MQ**) was given to each pupil, asking him/her to take it home, give it to his/her

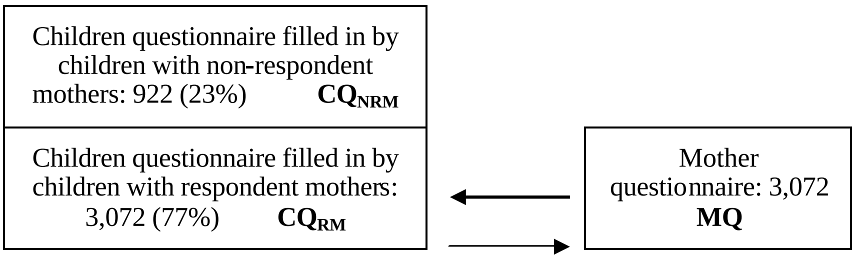
[2] UFS was part of a wider research programme founded by the Italian Ministry of Education and Research, directed by Massimo Livi Bacci, regarding several aspects of fertility of Italian women, who during the last fifteen years have been characterized by the lowest fertility in the world. The main results of this survey are reported in the special number of Genus (2004), edited by M. Livi Bacci, on the explanation of the Italian lowest-low fertility.

mother and bring it back to school. About 77% of mothers completed the **MQ**, with some differences among the four towns (see table 2). We can represent graphically (see figure 1) the structure of our data (for details see Giraldo, 2003).

Table 2: Sample size and response rates to the UFS in the four towns

	Udine	*Padua*	*Florence*	*Messina*	*Total*
Children questionnaire	814	792	696	1,692	3,994
Mothers questionnaire	544	627	480	1,421	3,072
% respondent mothers	*66.8*	*79.2*	*69.0*	*84.0*	*76.9*

Figure 1: Structure of the data of UFS



1.3. Three questions

The characteristics of women who did not answer to **MQ** can be explored using data collected by means of theirs sons with the children-questionnaire, i.e. $\mathbf{CQ}_{\mathbf{NRM}}$. The **CQ** can then be used to re-arrange **MQ** information, in order to take into account unit non-responses. Our analysis try to answer the following questions:

1) Can we trust children? And more generally, what is the *agreement* of the answers of children and their mothers?
2) Are responding mothers (and their families) different from non-responding mothers? If differences exist, are they relevant for our study? In other words, can these differences affect the results we draw from our data, i.e., the determinants of fertility?
3) If these differences are relevant for our study, i.e., unit non-responses induce bias, is there any method to correct the bias using information collected on children of responding and non-responding mothers?

The paper proceeds as follows. Section 2 describes procedures and methods used for answering the three questions, section 3 is devoted to results, and section 4 discusses some implications of results for social research.

2. Methods

2.1. Accuracy of the answers of children

In order to use **CQ** data to compare the two groups of responding and non-responding mothers, we need children to be good respondents, i.e., to report correctly some characteristics of their mothers[3]. We assume that each respondent mother correctly answers to the questions of her questionnaire. Even if questions here considered could seem very simple (number of children, education, citizenship…) this assumption, generally speaking, is not trivial, because self-filled questionnaires

[3] The response behaviour of children is investigate in Scott (1997).

are used. For example, in UFS survey some foreigner and/or low educated respondent mothers had serious problems in understanding questions written in Italian, in spite of our efforts of simplifying the questions. Their knowledge of Italian language is often worse than the one of their children. Consequently, in some cases it is possible that accuracy of **MQ** is lower than accuracy of **CQ**.

To study the response behaviour of children we compare the answers of children and mothers to the very same questions to verify their agreement. Since we have not answers of non-responding mothers, the agreement between the answers of mother and child is obviously studied only for the sub-sample of 3,072 couples of responding mothers and children (see table 2 and figure 1). Using this sub-sample we are implicitly assuming that children with responding and non-responding mothers are equivalent with respect to their degree of agreement between response at the two questionnaires. This is a necessary assumption: if the two groups were not equivalent in the knowledge of "families affaires", additional bias would be induced. The concordance between data collected by child and his/her mother is studied for five characteristics of mother: number of children, actual occupation (with or without a job), citizenship, education, age class.

The idea of using **CQ** to correct non-responses has been developed after the data collection: **CQ** was originally projected to collect some additional information on children consumptions, for a generic analysis of non-responses, for increasing the interest of pupil on the survey, in order to increase the mothers' response rate. Consequently, there is not a perfect correspondence between questions of **CQ** and **MQ**. This correspondence is very close for occupation, education and age class, whereas it is less satisfactory for the other two variables. Child was asked on their mother's *citizenship*, mother on her *place of birth*; child on *number of living brothers and sisters*, mother on number of *children ever born*. Consequently, for these two variables it is difficult to understand if discordance between $\mathbf{CQ_{RM}}$ and **MQ** is effective or rather due to difference between questions. Generally speaking, this procedure needs that the same questions are asked to the two components of the couple.

The agreement of responses given by children and mothers is measured by Cohen's k (Kappa) and weighted k. Simply stated, Cohen's k is an index of the proportion of agreement above chance, referring to a nominal classification. It is 0 when agreement equals that expected by chance, and is 1 when there is perfect agreement[4]. Weighted κ is appropriate when categories are numerous and ordered (e.g., for education), because in these circumstances the seriousness of any disagreement depends on the difference between ratings. Also for nominal politomous classifications, some disagreements may be considered more substantial than others. Weighted κ allows each cell i,j to be suitably weighted[5] according to the seriousness of disagreement between the i-th and j-th categories, for $i \neq j$. The asymptotic variance of Cohen's κ and that of weighted κ were computed according to Fleiss *et al.* (1969).

The effect of a high agreement rate (i.e. "we can trust children") is twofold: on one side we can contrast groups of families with responding and non-responding mothers, using information provided by the children. On the other side, data could be collected interviewing only pupils, avoiding to disturb their mothers, even if, obviously, determinants of fertility cannot be studied in their details without interviewing women. In this way, a tremendous simplification of data collection could be pursued, and non-response unit problem could be by-passed (as practically all the children present in the classroom filled in the questionnaire). The possibility of having good quality data interviewing only children is studied fitting some very simple logistic regression fertility models to $\mathbf{CQ_{RM}}$ (sub-sample of children with respondent mothers) and **QM**, where explanatory variables (referred to mother) are place of residence, citizenship, age, and education. Generally speaking, it could be possible that similar regression results are obtained even if concordance between answers of mother and child is not quite good.

[4] See, e.g., Agresti, 1990, pp. 366-367.

[5] We uses the weights $w_{ij} = 1 - |S_i - S_j| / (S_R - S_1)$, where S_i is the score for column i and R is the number of categories.

2.2. Comparing respondents and non-respondents mothers

As already stated, unit non-response in our survey reduces total sample size of about 23%, but what really matters is the possible bias induced by non-response. The question is: are responding and non-responding mothers different? If this were the case, results drawn from our sample can be far from those drawn from the complete sample, i.e., the sample we would have observed if all mothers had been interviewed. We can try to answer the question just looking at information derived by **CQ**. If there is an agreement in the response of children and mothers, we can characterize families of responding and non-responding mothers finding out differences between the two groups.

To this purpose, some descriptive and univariate analyses are performed on the data. Univariate distributions of variables collected by **CQ** are compared for respondents and non-respondent mothers. In order to consider also interactions between variables in determining the "response behavior" we model, using again the **CQ**, the probability of a mother to participate to the survey by means of a logistic regression and we look at the variables driving the non-response behavior of the mothers.

The main aim of UFS survey was to study the determinants of fertility. Consequently, we control if the association between explanatory variables and fertility is different for responding and non-responding women. Fertility is separately modeled in two groups: children with responding mother ($\mathbf{CQ_{RM}}$) and children with non-responding mother ($\mathbf{CQ_{NRM}}$). Two logistic regression models are fitted to each group of data: in the first one, women with one child contrast women with two children or more, in the second one women with at least two children contrast women with three children or more.

If results of a logistic regression from $\mathbf{CQ_{RM}}$ are close to the ones from $\mathbf{CQ_{NRM}}$, it means that fertility behaviour and its determinants are similar for responding and non-responding mothers. We can test formally this similarity constructing a test for verify the global identity of the coefficients of the two logistic regression. Moreover if we fit a logistic regression on the entire sample of children based on **CQ** and the results are similar to the ones of $\mathbf{CQ_{RM}}$, we can conclude that non-responses are not "strong" enough to disturb the fertility analysis, because (a) they are relatively few, or (b) association structure is similar for respondents and non-respondents women, or (c) these two reasons jointly hold. For each logistic regression, the same procedure of backward selection of explanatory variables is used: we start from models with rather all the variables collected in **CQ**, preserving variables that are statistically significant for at least one model. Models here presented – even if they are not satisfactory in order to "explain" fertility behavior – are rich enough to study differences of the regression structure in the groups.

2.3. A methodology for correcting unit non-response in MQ using CQ

Results of the above analysis can lead us to conclude that the two groups of responding and non-responding mothers are different with respect to their fertility behavior, as well as their attitude on responding. If this is the case, conclusion coming from the sub-sample of responding mothers can be far from the situation we could have found if all mothers of the contacted children had been interviewed.

As stated above, in cross sectional survey without any other supplemental information it is difficult to prove that bias induced by unit non-response exists and it is even more difficult to correct for it. In our survey, data collected by **CQ** give us some supplemental information, that can be used to construct a weighting adjustment to compensate for the potential bias induced by unit non-response. **MQ** can be seen as the second wave of a panel survey (Kasprzyk *et al.*, 1989), whose first wave is **CQ**, and unit non-responses $\mathbf{CQ_{NRM}}$ are the units lost by attrition.

Consequently, correction procedures used to manage the effect of attrition in panel survey can be applied (see, for example, Kalton, 1986). The rationale is to "weight" respondent women to represent also non-respondent women (Holt and Elliot, 1991). There are several weighting procedure that can be applied to correct for attrition in panel survey. We choose the one based on *propensity scores* (David *et al*., 1983; Little, 1986; Giraldo *et al*., 2001). A propensity score (Rosenbaum and Rubin, 1983; Rosenbaum, 1995) in observationally studies is the conditional probability of exposure to treatment rather than control given observed covariates, or, more generally, the conditional probability of selection given observed covariates. It may be used to adjust for nonrandom treatment assignment or nonrandom selection. It may be estimated, for example, by means of a logistic regression (Hosmer and Lemeshow, 1989) on data known for all the sample (in our case, data collected by **CQ**), where the dichotomous variable modeled is participation – non-participation. The inverse of the probability of a mother to participate to the survey, conditional on a set of explanatory variables (age, number of children, citizenship, …), estimated by a logistic regression, can then be used as weight in order to compensate bias induced by non-participation.

3. Results

3.1. A good concordance between answers of mothers and children

The answer of children and mothers are not dissimilar as regard to the five questions in common in **CQ** and **MQ** (see table 3 – 5). The concordance is very good for *number of children*, less for *age*, *education* and – particularly – *employment status*[6].

*Table 3: Comparison between distribution of number of children, collected by means of **MQ** (respondent mother) and **CQ**$_{RM}$ (sub-sample of children with respondent mother).*

	***MQ** Mothers respondent at home*					
***QC**$_{RM}$ (sub-sample of children with respondent mother)*	1	2	3	4 or more	Missing	Total
1	*470*	22	2	0	5	499
2	40	*1,632*	35	9	18	1,734
3	7	44	*554*	15	14	634
4 or more	5	13	17	*167*	4	206
Total	522	1.711	608	191	41	3,073

*Table 4: Comparison between distribution of mother's employment status, collected by means of **MQ** (respondent mothers) and **CQ**$_{RM}$ (sub-sample of children with respondent mother).*

	***MQ** Mothers respondent at home*		
***CQ**$_{RM}$ (sub-sample of children with respondent mother)*	Mother works	Mother doesn't work	Total
Mother works	*1,772*	101	1,873
Mother doesn't work	138	*524*	662
Total	1,910	625	2,535

Table 5: Comparison information child-mother: simple and weighted Coehn's Kappa

Variable	*Simple Kappa*	*Weighted Kappa*	*Confidence limits (95%)*	
Number of children	-	0,8977	0,8823	0,9131
Employment status	0,7512	-	0,7214	0,7809
Education	-	0,7774	0,7590	0,7959
Age (in classes)	-	0,8028	0,7842	0,8214

As table 6 shows, these differences do not influence the association between fertility and the explanatory variables, as the odd ratio of each category is practically the same for logistic models fitted on questionnaires filled in by mothers or by children of respondent mothers.

[6] This variable has a certain number of item non-response, either for mothers and children. For this the reason the totals of table 3 and 4 are different.

Consequently, using data collected in the classrooms it could be possible to obtain the same results for analysing both simple frequencies of our five variables and association between fertility and some explanatory variables. Of course, it does not mean that it is possible to study determinants of fertility simply interviewing children: they are able to report only simple information on their mothers. However, this strong concordance between responses of children and mother is an indispensable pre-condition in order to use data collected interviewing pupils to study non-responses, and correcting their influences in fertility analysis.

*Table 6: Comparison between logistic analysis of fertility obtained using questionnaires of mothers, **MQ** and children of respondent mothers, $\mathbf{CQ_{RM}}$. Odds ratio for the two groups. For each explanatory variable, the reference category (odds ratio = 1.000) is in brackets.*

	Probability to have one child vs. two children or more		*Probability to have two children vs. three children or more*	
	MQ *Responding mothers*	$\mathbf{CQ_{RM}}$ *Children with responding mother*	***MQ*** *Responding mothers*	$\mathbf{CQ_{RM}}$ *Children with responding mother*
Place of residence (Udine)				
Padua	0.735**	0.750*	0.797	0.841
Florence	1.094	1.094	1.064	1.238
Messina	0.271***	0.273***	0.420***	0.465***
Citizenship of mother (Italian)				
Foreign	1.233	1.234	0.815	0.829
Mother's age (<37)				
37-44	0.739	0.846	0.721**	0.689***
More than 44	0.629**	0.664**	0.430***	0.445***
Mother's education (low)				
High school	1.341**	1.432***	1.945***	1.682***
University	1.426**	1.604***	2.518***	2.383***
Sample size	2,868	2,780	2,885	2,780
Deviance (a)		148.420(8)	166.501(8)	158,554(8)

*** p<0.01 ** 0.01<p<0.05 * 0.05<p<0.10
(a) Deviance = -2 (log-likelihood of the model with constant term only - log-likelihood of the actual model). Degrees of freedom in brackets.

3.2. Differences between responding and non-responding mothers

Response probability changes with some characteristics of children and their mothers (table 7). First of all, response rate is higher in Messina and Padua, lower in Florence and Udine (return also to table 2). It is likely that in Messina the response rate was higher than elsewhere because children bringing back **MQ** received a little gift.

Let us consider some characteristics of the children. Response rate is higher among mothers of regular students (i.e. pupils who did never lose one year of school or more), mothers of female students, mothers of students who do not know their father's education. It could be that male students and pupil less involved in school and educational matters are less diligent in following the indications of interviewers.

Moreover, response rate is influenced by some characteristics of the mothers. Response rate is lower for foreigner mothers, more educated women (in the North) and less educated ones (in Messina), mother of child(ren) having a Play Station, mothers of three children or more (see also table 8). The first of these results were expected, whereas the meaning of the other results are more ambiguous. Perhaps, mothers with many children had less time for responding a questionnaire that asked additional time to the more fertile women.

In spite of motivations leading the behaviors of children and mothers, it is important to underline that respondents are different from non-respondents, for both fertility and some characteristics that could be associated to fertility. Consequently, respondents are a non representative sub-sample of the whole sample and fertility analysis on **QM** could be biased.

Table 7: Logistic regression for response probability, children questionnaire. Odds ratios for the four towns (§). For each explanatory variable, the reference category is in brackets (odds ratio = 1.000).

Variables	*Udine*	*Padua*	*Florence*	*Messina*	*All towns*
Place of residence (Udine)					
Padua	-	-	-	-	1.906***
Florence	-	-	-	-	1.162
Messina	-	-	-	-	3.048***
Regular student (yes)					
No	0.600**	0.488**	0.510**	0.894	0.631***
Sex (Male)					
Female	1.009	1.352	1.279	1.320*	1.222**
Citizenship of mother (Italian)					
Foreign	0.966	0.630	0.711	0.691	0.724**
Father education (low)					
High school	0.867	1.091	0.940	1.094	1.015
University	0.833	0.651	0.535**	1.544**	0.845
Don't know	0.647	0.684	0.405***	0.892	0.681***
N° of children (2 or less)					
3 or more	0.649**	0.716	0.548***	0.640***	0.636***
Play station (no)					
Yes	0.645***	0.585***	0.701*	0.718**	0.672***
Personal computer (no)					
Yes	1.685**	1.219	1.050	1.102	1.276**
Sample size	804	769	677	1,655	3,905
Deviance (a)	34.1 [9]	32.6[9]	49.0[9]	32.9 [9]	240.5 [12]

*** $p<0.01$ ** $0.01<p<0.05$ * $0.05<p<0.10$

(a) Deviance = -2 (log-likelihood of the model with constant term only - log-likelihood of the actual model). Degrees of freedom in brackets.

(§) A logistic regression procedure was performed separately for each town. For each town, the same procedure of backward selection of explanatory variables is used, starting from models with almost all the variables collected in **CQ**. In order to compare results for different towns, every variable that is statistically significant for at least one town (or in the all towns model) is included in the models.

In order to verify this possibility, in tables 9 and 10, we compare some explanatory fertility models applied to all the children (**CQ**), children with responding mother (**CQ**$_{\mathbf{RM}}$), and children with non-responding mother (**CQ**$_{\mathbf{NRM}}$). The association structure deducted by **CQ**$_{\mathbf{NRM}}$ is quite different from the one suggested by **CQ**$_{\mathbf{RM}}$ (compare the last two columns of the two tables). For example the influence of citizenship on fertility is clearly stronger among non-responding women (but not significant); similar results hold for education, age of the parents and other explanatory variables.

Luckily, in our specific situation, these differences are not strong enough to bias the analysis of fertility determinants. In fact, odds ratios calculated using questionnaires compiled by children with non-responding mothers are the same as the ones calculated using questionnaires of children with responding mother (compare the last two columns of the two tables). We test formally this similarity constructing a test for verify the global identity of the coefficients of the two logistic regression. What we found is that the coefficients of the logistic model for children with responding and non-responding mother are statistically equal. As far as this model of fertility is concerned, in our sample, there are not differences among the two groups.

Table 8: Non-response rate by number of children of mother. CQ in the four towns.

Number of children	*Udine*	*Padua*	*Florence*	*Messina*	*All towns*
1	30%	21%	24%	15%	23%
2	40%	19%	29%	13%	22%
3+	41%	25%	43%	20%	28%
Total	33%	21%	31%	16%	23%

Table 9: Probability to have one child vs. two children or more. Odds ratios for three groups, children questionnaire. For each explanatory variable, the reference category (odds ratio = 1.000) is in brackets

	All the children CQ	*Children with responding mother* ***CQ_{RM}***	*Children with non-responding mother* ***CQ_{NRM}***
Place of residence (Udine)			
Padua	0.819	0.766*	0.965
Florence	1.032	1.083	0.876
Messina	0.310***	0.295***	0.334***
Citizenship of mother (Italian)			
Foreign	0.822	1.038	0.556*
Age of father (<41)			
41-48	0.919	0.894	1.041
More than 48	0.679***	0.640***	0.856
Don’t know	1.259	1.107	1.846*
Mother education (low)			
High school	1.446***	1.339**	1.949**
University	1.411**	1.351*	1.691*
Don’t know	1.166	1.007	1.914*
Mobile (no)			
Yes	1.681***	1.649***	1.747**
Play station (no)			
Yes	0.710***	0.710***	0.761
At least a travel by plane (no)			
Yes	1.501***	1.551***	1.356
Sample size	*3,906*	*3,012*	*894*
Deviance (a)	234.124 [13]	199.691 [13]	47.883 [13]

*** $p<0.01$ ** $0.01<p<0.05$ * $0.05<p<0.10$

(a) Deviance = -2 (log-likelihood of the model with constant term only - log-likelihood of the actual model). Degrees of freedom in brackets.

Table 10: Probability to have two children vs. three children or more. Odds ratio for three groups, children questionnaire. For each explanatory variable, the reference category (odds ratio = 1.000) is in brackets.

	All the children **CQ**	Children with responding mother **CQ**$_{RM}$	Children with non-responding mother **CQ**$_{NRM}$
Place of residence (Udine)			
Padua	0.910	0.820	0.937
Florence	1.102	1.147	0.980
Messina	0.482***	0.419***	0.466***
Citizenship of mother (Italian)			
Foreign	0.558***	0.734	0.479***
Age of mother (<37)			
37-44	0.903	0.916	0.863
More than 44	0.599***	0.563***	0.709
Don't know	0.421***	0.358***	0.559
Age of father (<41)			
41-48	0.860	0.760*	1.104
More than 48	0.743**	0.640**	1.042
Don't know	1.627*	1.781	1.697
Mother education (low)			
High school	1.419***	1.442***	1.338
University	1.300**	1.385**	1.170
Don't know	1.149	1.273	1.049
Father education (low)			
High school	1.497***	1.566***	1.230
University	2.112***	2.247***	1.671**
Don't know	1.187	1.263	1.061
Mobile (no)			
Yes	1.431***	1.558***	1.133
Play station (no)			
Yes	0.918	0.829**	1.380**
Many books at home (no)			
Yes	0.798***	0.721***	0.928
Sample size	3,919	3,026	893
Deviance (a)	256.285 [19]	242.948 [19]	51.042 [19]

*** p<0.01 ** 0.01<p<0.05 * 0.05<p<0.10

(a) Deviance = -2 (log-likelihood of the model with constant term only - log-likelihood of the actual model). Degrees of freedom in brackets.

3.3. Correcting for unit non-response

We apply the weighting methodology described sub 2.3. **MQ** data are weighted by means of the inverse of propensity score (the inverse of the probability of a mother to respond at **MQ** as estimated in table 7) and the distribution of some variables weighted and not weighted are compared. We present in table 11 and 12 the results for the number of children and citizenship. These distribution using weighted and not weighted data are quite similar. These results confirm what we found in the previous paragraph: in this specific case, **MQ** data can be used to study the determinants of fertility without too many problems, because the missing data do not influence the association between fertility and explanatory variables.

In general, when we estimate models conditioned to observed variables, the fertility models above are an example, there is no need to use weighted data as the conditioning to observed covariates correct the bias induced by non-response. In case of non conditional analysis, as univariate distributions or means, it is important to use weights if we are aware of some sort of bias induced by non-response.

*Table 11: Distribution of number of children (%), collected by means of **MQ**, weighted or not with inverse of propensity score.*

	MQ	
Number of children	*Not weighted*	*Weighted*
1	17.22	17.38
2	56.43	55.00
3	20.05	21.08
4 or more	6.30	6.53
Total	100.00	100.00

*Table 12: Distribution of citizenship (%), collected by means of **MQ**, weighted or not with inverse of propensity score.*

	MQ	
Citizenship	*Not weighted*	*Weighted*
Italian	95.58	94.59
Foreign	4.42	5.41
Total	100.00	100.00

4. Conclusions

We propose a combined procedure of data-collection, non-response analysis, and non-response correction, in order to control the possible bias induced by unit non-response in social surveys, in countries where primary education is practically universal. The idea is to “simulate” a panel procedure, collecting few data in the classrooms among a random sample of pupils (first wave) and more extensive data among their parents or relatives (second wave), “using” children as messengers for questionnaires to be filled in at home by their parent(s) or relatives. Unit non-response can be considered as attrition in the panel surveys, and consequently methods proposed to deal with attrition can be used, in order to manage non-response, obtaining results representative for the whole population.

In our opinion, the merits of this procedure are the following:

(1) if some questions are identical in the two questionnaires, it is possible to control the concordance between answers of children and his/her relative;

(2) if some characteristics of relatives are asked to pupils, it is easy to control if responding and non-responding relatives are similar;

(3) if respondents and not-respondents are not similar and non-response rate is high, it is possible to adjust data of respondents, using weights obtained managing data collected among pupils. With this procedure, adjusted results become statistically representative not only for the (unknown) population sampled by respondents, but for all the population, even if respondent rate is low, and characteristics of respondents and non-respondents are quite different.

It is clear that this procedure can be used only if the target population of the survey overlap relatives of pupils (e.g. fathers or mother of defined age and parity condition). This method of contacting adult population can drive a “strange” sample structure, that has to be carefully taken into account. In other words, we should always remember that we sample pupils rather than their correspondent relatives. For example, in UFS the probability of a woman to be interviewed grows with her parity: it is zero for women without children, **p** for women with one child, 2**p** for women with two children, and so on. In this case, in order to “transform” a sample of pupils into a sample of mothers, each mother should be weighted by 1/**k**, where **k** is the number of her children.

Another problem is that if self-filled questionnaires are used, it is not possible to control misunderstanding of questions, and the structure of questionnaire has to be very simple (e.g. it is not possible to use complex filtering questions). However, in our opinion, the qualities of this cheap data-collection procedure suggest to explore its possible refinements, getting over its limits. For example, it could be possible, when submitting questionnaire to pupils, to ask them phone number or address of their parent(s) or relative; in a second wave, relatives could be interviewed face-to-face or using a CATI procedure. If only non sensitive questions are asked to pupils, this procedure could avoid privacy problems.

Up to now, we have described a survey structure based on data collected among pupils and their relatives. However, this procedure to control non-responses can be generalized: what is important is collecting for all the sample units some data hypothetically related to the propensity of non-response, in order to post-stratificate the sample. If the list comes from population register, it could be enough to select some data already available (e.g. sex, age, marital status, number of children living with mother/father, and so on). In a CATI survey, it could be enough to ask the person answering the phone few basic data on the person eligible for the survey (e.g. in a fertility survey: sex, age, marital status, number of children, education), before asking if eligible person wants to answer the complete questionnaire. In a CATI survey, however, using this procedure it is not possible to control the non-response if no member of the eligible person's family is contacted by phone. It could be a problem if the probability of telephonic contact is associated with some characteristics collected in the survey.

Unit non-response problems should not be unevaluated. In the case of UFS we are lucky, as association structure for explaining fertility is the same when non-respondents are or are not taken into account – thanks to a relatively low non-response rate (23%) – even if responding and non-responding mothers are clearly different but do not affect fertility and its determinants. Generally speaking, without collecting supplementary data for all the target population, we can never be sure that non-response could influence results, and results obtained from only respondents could give a biased representation of reality.

References

Agresti A. (1990) *Categorical data analysis*, Wiley, New York.

Banca d'Italia (2000) "I bilanci delle famiglie italiane nell'anno 1998", *Supplemento al Bollettino Statistico* (nuova serie), Roma, Banca d'Italia.

Banca d'Italia (2002) "I bilanci delle famiglie italiane nell'anno 2000", *Supplemento al Bollettino Statistico* (nuova serie), Roma, Banca d'Italia.

Fleiss J.L., J. Cohen and B.S. Everitt (1969) "Large-sample standard errors of Kappa and Weighted Kappa", *Psychological Bulletin*, 72, 323-327.

Dalla Zuanna G., Salvini S. (2001) "L'indagine "Troppi o nessuno": dalle ipotesi alle possibili verifiche", *Workshop "La bassa fecondità italiana tra costrizioni economiche e cambio di valori"*, Dipartimento Statistico, Firenze 8-9 Novembre 2001.

Dalla Zuanna G., Salvini S. (2003) "Un quadro generale della fecondità nelle realtà urbane", in M. Breschi e M. Livi Bacci (a cura di), *La bassa fecondità italiana tra costrizioni economiche e cambio di valori. Presentazione delle indagini e dei risultati*, Forum, Udine, 77-106.

David M., Little R.J.A., Samuhel M., Triest R. (1983) "Nonrandom nonresponse models based on the propensity to respond", *Proceedings of the Business and Economic Section*, American Statistical Association, 168-173.

De Sandre P., Ongaro F., Rettaroli R., Salvini S. (1997) *Matrimonio e figli: fra rinvio e rinuncia*, il Mulino, Bologna.

Festy P., Prioux F. (2002) *En evaluation of family and survey project*, United Nations, New York and Geneva.

Giraldo A., E. Rettore, U. Trivellato, (2001) "Attrition bias in the Bank of Italy's SHIW", *Proceedings of the International Conference on Quality in Official Statistics*, Stockholm, May 14-15, 2001 and *Working paper n. 41*, Gruppo di ricerca: "Occupazione e disoccupazione in Italia: misura ed analisi dei comportamenti".

Giraldo A., (2003) "Qualità dei dati nei questionari 'Donne' e 'Bambini'", in M. Breschi e M. Livi Bacci (a cura di), *La bassa fecondità italiana tra costrizioni economiche e cambio di valori. Presentazione delle indagini e dei risultati*, Forum, Udine, 43-56.

Kalton G., (1986) “Handling Wave Nonresponse in Panel Surveys”, *Journal of Official Statistics*, 2, 303-314.

Kasprzyk, D., Duncan, G.J., Kalton, G. and Singh, M.P. (Eds.) (1989) *Panel surveys*, Wiley, New York.

Holt D., Elliot D., (1991) “Methods of weighting for unit non response”, *The Statistician*, 40, 333-342.

Hosmer D. W., Lemeshow S. (1989) *Applied Logistic Regression*, Wiley, New York.

Istat (2002), *I consumi delle famiglie*, Anno 2001, (www.istat.it).

Little R.J.A., (1986), “Survey nonresponse adjustment for the estimates of means”, *International Statistical Review*, 54, 139-157.

Little R.J.A, Rubin D.B. (2002) *Statistical analysis with missing data*, 2nd ed., Wiley, New York.

Scott J. (1997) “Children as respondent: methods for improving data quality”, in L. Lyberg et al. (eds.) *Survey Measurement and Survey Quality*, Wiley, New York, 331-350.

Rosenbaum P.R. (1995) *Observational studies*, Springer-Verlag, New York.

Rosenbaum P.R., Rubin D.B., (1983) “The central role of the propensity score in observational studies for causal effects”, *Biometrika*, 70, 41-55.

Acknowledgements

We are would like to thank Enrico Rettore for helpful comments and the participants at the Workshop on Item-Non-response and Data Quality in Large Social Survey, Basel, Switzerland, October 9-11, 2003.

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order copies of the working papers from by emailing to wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it