Non-Sampling Errors in Household Surveys: The Bank of Italy's Experience

Giovanni D'Alessio and Giuseppe Ilardi

Abstract Non-sampling errors are a serious problem in household surveys. This paper exploits the Bank of Italy's Survey on Household Income and Wealth to show how these issues can be studied and how the main effects on estimates can be accounted for. The topics examined are unit non-response, uncorrelated measurement errors and some specific cases of underreporting. The unit non-response can be overcome by weighting valid cases using external (typically demographic and geographical) information or by modelling the respondents' propensities to participate in the survey. The effect of the uncorrelated measurement errors can be evaluated using specific reliability indices constructed with the information collected over the panel component. The underreporting bias of income and wealth is estimated by combining statistical matching techniques with auxiliary information and by exploiting different response behaviours across different groups.

1 Introduction

Errors in survey data can be divided depending on the source into two broad categories: sampling and non-sampling errors. The former includes errors in estimating the relevant population parameters derived from the inferential process: these tend to vanish as the sample size increases. Non-sampling errors mainly relate to measurement design, data collection and processing.

Non-sampling errors comprise quite diverse specific types of error that are usually harder to control than sampling ones. Following Biemer and Lyberg (2003), we can classify the non-sampling errors as: specification error; coverage or frame error;

G. D'Alessio (✉) and G. Ilardi
Bank of Italy, Economic and Financial Statistics Department, Rome, Italy
e-mail: giovanni.dalessio@bancaditalia.it

G. Ilardi
e-mail: giuseppe.ilardi@bancaditalia.it

C. Davino and L. Fabbris (eds.), *Survey Data Collection and Integration*,
DOI: 10.1007/978-3-642-21308-3_7, © Springer-Verlag Berlin Heidelberg 2013

processing error; unit non-response; and measurement errors.[1] Usually non-sampling errors affect both bias and the variance of estimators; and their effects do not necessarily diminish as sample size increases. In many economic applications, the non-sampling component of total error outweighs the sampling one.[2] This is the case in many of the variables collected in the Bank of Italy's Survey of Household Income and Wealth (SHIW). The survey estimate of total household net wealth is approximately half the corresponding value deriving from the financial accounts (FA). True, the FA data rely on many measurement hypotheses and are subject to errors; nevertheless this discrepancy cannot be attributed to sample variability and is likely to depend on non-sampling errors—presumably because of a lower propensity of wealthier households to participate in the survey and/or widespread underreporting by respondents of their assets. This evidence is the Bank of Italy's strongest motivation for its efforts to analyse non-sampling errors for the household budget survey. In the next sections we evaluate non-sampling errors that typically occur in the SHIW. This informal approach allows the discussion of some of the typical problems associated with using household data.[3]

After a brief description of the SHIW (Sect. 2), we describe the survey experiences with non-response (Sect. 3.1), measurement errors (Sect. 3.2) and underreporting (Sect. 3.3). Section 4 concludes.

2 The Survey on Household Income and Wealth

Since 1965, the SHIW gathers data on Italian households' income, wealth, consumption and use of payment instruments. It was conducted annually until 1984 and biannually since (with the exception of 1998). The sample consists of about 8,000 households (secondary units) in 350 municipalities (primary units), drawn from a population of approximately 24 million households. The primary units are stratified by region and municipality size. Within each stratum, the selected municipalities include all those with a population of more than 40,000 units (self-representing municipalities), while the smaller towns are selected with probability proportional

[1] A specification error occurs when the collected data do not include relevant economic variables for the survey objectives. A coverage error exists when some statistical units belonging to the reference population are not included in the sampling frame. Non-response errors occur because some households do not participate in the survey. Measurement errors arise during the data collection process; errors made by the interviewer or by the respondent, and the mode of data collection contribute to measurement error. Processing errors include errors emerging from data entry, computer programs (i.e. miscalculation of the weights) or incomplete instructions. An alternative classification distinguishes non-sampling errors on the base of the source of such errors; for instance, the interviewer may affect both unit non-response, item non-response and measurement errors (Blom 2011).

[2] In budgeting a survey there is a clear trade-off between the two types of error. Resources can be devoted to procuring a large sample and thus minimizing random sampling errors or else concentrated on a smaller sample but with better interviewer controls, a higher response rate and more accurate data collection procedures.

[3] See Lessler and Kalsbeek (1992) for a general exposition on non-sampling errors.

to size. At the second stage, the individual households are selected randomly from the population register.[4,5] Through 1987 the survey used time-independent samples (cross sections) of households. In order to facilitate the analysis of changes, the 1989 survey introduced a panel component, and almost half of the sample now consists of households interviewed in one or more previous waves. Data are collected by a market research institute through computer-assisted personal interviews. Households answer an electronic questionnaire—that not only stores data but also performs a number of checks so that data inconsistencies can be remedied directly in the presence of the respondent. The Bank of Italy publishes a regular report with the main results, the text of the questionnaire and the main methodological choices. Anonymized microdata and full documentation can be accessed online for research purposes only (microdata are available from 1977 onwards). Recent economic studies based on this survey have covered such topics as households' real and financial assets over time; risk aversion, wealth and financial market imperfections; dynamics of wealth accumulation; payment instruments used; and tax evasion. The financial section has been extensively exploited for studies on the financial structure of the Italian economy. The SHIW is also part of the European household survey promoted by the euro-area national central banks in order to gather harmonized data on income and wealth.

3 Unit Non-Response and Measurement Errors in the SHIW: Some Empirical Studies

3.1 The Analysis of Unit Non-Response

In most household surveys not all the units selected will participate. The difference between the intended and the actual sample reflects both unwillingness to participate (refusals) and other reasons (most commonly, "not at home"). This may have serious consequences for survey statistics, which need to be properly addressed. Let us consider the case of units that are selected to be surveyed but do not participate. Denoting by y_r the values of variable y for the group of n_r respondents and by y_{nr} the values for the unobserved group of $n - n_r$ non-respondents, the estimator of the mean can be decomposed into two parts

$$\bar{y} = \frac{n_r}{n}\bar{y}_r + \frac{n - n_r}{n}\bar{y}_{nr}. \tag{1}$$

[4] Since households are extracted from the registry lists, the reference population does not include Italian citizens living in institutions (prisons, barracks, nursing homes or convents).

[5] Respondents receive a participation letter explaining the purpose of the survey, a booklet describing the main uses of the information and a small gift; a toll-free telephone number is available to supply any information about the survey.

The expected value of $\bar{y}$ is given by $\mu = f\mu_r + (1-f)\mu_{nr}$, where f is the response rate, i.e. the share of responding units in the population, and μ_r and μ_{nr} are the population means of the responding and non-responding units respectively.

The estimator computed on respondents only, $\bar{y}_r$, is a biased estimator of μ, with a bias given by

$$E(\bar{y}_r) - \mu = (1-f)(\mu_r - \mu_{nr}). \tag{2}$$

The magnitude of non-response bias depends both on the non-response rate $1-f$ and on the difference between μ_r and μ_{nr}. When non-response occurs, the estimator $\bar{y}_r$ will be biased unless the pattern of non-response is random, that is the assumption $\mu_r = \mu_{nr}$ holds.

In household surveys, however, we cannot assume that non-responses are totally random; both the sample units that refuse to participate and those that are not at home tend to belong to specific population groups; so we need a procedure to correct for the bias.[6]

If we knew the participation probability p_i of household i, an unbiased estimator of the population mean could be obtained by extending the Horvitz–Thompson estimator (Little and Rubin 1987)

$$\bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}, \tag{3}$$

where $w_i = 1/(\pi_i p_i)$, to include both the probability of being included in the sample π_i and the probability of actually participating p_i.[7] We assume that these two sets of weighting coefficients are independent of each other. In order to correct for non-response, we need information on the selection process governing the response behaviour. But how can we obtain information on this process, given that non-respondents—by definition—are not reached by interviewers or deliberately avoid participation?

Several statistical techniques, based on various assumptions, can be employed. Knowledge of the distribution of some relevant characteristics for the entire population allows us to compare the sample with the corresponding census data. A significant deviation of the sample distribution from that of the population gives us indirect information on the selection process. The sample composition can thus be aligned with the population distributions by means of post-stratification techniques.[8]

[6] See Särndal and Lundström (2005) for a recent review of estimation methods to account for non-response.

[7] Many practitioners believe that the purpose of weighting is to reduce non-response bias, at the cost of increasing the variance of the estimates and transforming the efficacy of weighting adjustments into a bias-variance trade-off. However, Little and Vartivarian (2005) point out that if the weighting adjustments are positively correlated with the survey outcome, then the weighting system can also reduce sampling variance of the estimates.

[8] When only marginals are known, the technique employed is called as Iterative Proportional Fitting or Raking (Kalton and Flores Cervantes 2003). More in general, the calibration techniques, based on the linear regression model, offer a wide variety of solutions in adjusting the sample weights so

The SHIW data show a higher frequency of elderly persons than the census of the population, while younger persons are underrepresented. Post-stratification is a common practice of embedding into estimators information about population structure; the procedure can also reduce the variability of the estimates. Unfortunately, the information available for post-stratification is often limited (sex, age, education, region, town size) and as such is insufficient for a complete detection of non-response behaviour.

As a part of the SHIW sample consists of households already interviewed in past waves (the panel component), information on the propensity to participate can be obtained by an analysis of attrition, i.e. non-participation of a panel household in a subsequent wave of the survey. Following this approach, Cannari and D'Alessio (1993) found that non-response characterizes households in urban areas and in the northern Italy; and that participation rates decline as income rises and household size decreases. The relationship with the age of the head of household is more ambiguous: not-at-homes decline sharply with age but refusals and other forms of non-participation increase. On the basis of these findings, Cannari and D'Alessio estimated that non-participation caused a 5.4 % underestimate of household income in 1989.

This approach cannot be considered fully satisfactory; in fact, its validity depends on the assumption that the pattern of attrition within the panel component can be assimilated to non-participation of households contacted for the first time. Actually, a household's decision to participate in the survey may have been influenced by a previous interview and the estimation of the attrition pattern can shed light only on some aspects of non-response.

In many cases, some characteristics of non-respondents can be detected. In conducting personal interviews, for example, the characteristics of the neighbourhood and of the building are observable. In the most recent SHIW waves, several sorts of information on non-respondents have been gathered. Comparing respondents and non-respondents as regards these characteristics can help us understand the possible bias arising from non-response.

Information on the characteristics of non-respondents can also be inferred by analyzing the effort required to get the interview from responding households. The survey report usually includes a table with the number of contacts needed to obtain an interview, according to the characteristics of the households. In 2008, in order to get 7,977 interviews a total of 14,839 contacts was attempted (Banca d'Italia 2010b).[9] The difficulty of obtaining an interview increased with income, wealth and the educational attainment of the household head. It was easier to get interviews

(Footnote 8 continued)
as to reproduce ancillary external known information. Singh and Mohl (1996) provide a detailed description of many of these methods.

[9] The households that could not be interviewed were replaced by others selected randomly in the same municipality.

in smaller municipalities, with smaller households and with households headed by retired persons or women.[10]

We can compare the households interviewed at first visit with those that have been interviewed after both an initial refusal or a failure in the contact (not at home). These two groups offer valuable information on non-response. The households successfully interviewed after first being found not-at-home and that who initially refused to participate appear to have a higher income and wealth than the average sample (for the two groups, by 5.0 and 21.6 % for income and by 5.5 and 27.1 % for wealth respectively).

Assuming that the households interviewed after an initial not-at-home or after a refusal can provide useful information on non-responding units, we can estimate the bias due to non-response. An adjusted estimate can be obtained by re-weighting the interviewed households by the inverse of their propensity to participate. The results for the 1998 survey (D'Alessio and Faiella 2002) showed that wealthier households had a lower propensity to participate in the SHIW. Thus the adjusted estimates of income and wealth are higher than the unadjusted estimates. The correction is smaller for income and for real wealth, more significant for financial assets (ranging respectively from 7 to 14 %, 8 to 21 % and 15 to 31 %, depending on the model adopted).[11]

Different estimates of the effects of unit non-response on sample estimates were obtained by a specific experiment carried out in the 1998 survey. A supplementary sample of about 2,000 households, customers of a leading commercial bank, was contacted, 513 of which were actually interviewed.[12] For these out-of-sample households, the SHIW gathered data on actual financial assets held, the results of the current and supplementary samples were similar.[13]

3.2 Measurement Errors: Uncorrelated Errors

One of the most important sources of error in sample surveys is the discrepancy between the recorded and the "true" micro-data. These inconsistencies may be due to response errors or to oversights in the processing phase prior to estimation.

[10] In the most recent wave, the Bank of Italy conducted an experiment aiming to evaluate the effect of the gift on the participation.

[11] D'Alessio and Faiella's method belongs to the class of sequential weight adjustment (Groves and Couper 1998; Iannacchione 2003) which constructs the non-response adjustment weights by modelling the information on the two-stage response process, contact and participation. A different class of non-response adjustment that can be used in a regression analysis is the sample selection models (Heckman 1979). In this framework, the economic relation of interest is modelled with an additional regression equation that account for the censoring of non-participating households. In this strand of literature, a recent work by De Luca and Peracchi (2011) proposes an adjustment procedure both for the item and the unit non-response using semiparametric inference.

[12] The supplementary sample was drawn from a list of clients following a stratified random sample method, with a higher sampling rate for wealthier households.

[13] A strict protocol was devised to guarantee full protection of the respondents' confidentiality.

Irrespective of the reasons, the effects of errors on estimates are seldom negligible, so we need to evaluate their size and causes.

Involuntary errors in reporting values of some phenomena (e.g. the size of one's dwellings), due to rounding or to lack of precise knowledge, may still cause serious problems for estimators.

Consider a continuous variable X measured with an additive error: $Y = X + \varepsilon$. The measure Y differs from the true value X by a random component with the following properties: $E(\varepsilon) = 0$; $E(X, \varepsilon) = \sigma_{X,\varepsilon} = 0$; $E(\varepsilon^2) = \sigma_\varepsilon^2$. This type of disturbance is called homoscedastic and with uncorrelated measurement error. Under these assumptions, the average of Y is an unbiased estimator of the unobservable variable X–as $E(Y) = E(X)$–while the variance of Y is a biased estimator of the variance of X. In fact

$$\sigma_Y^2 = \sigma_X^2 + \sigma_\varepsilon^2 = \frac{\sigma_X^2}{\lambda^2}, \tag{4}$$

where $\lambda^2 = \sigma_X^2/\sigma_Y^2$ is the reliability coefficient. Therefore, the index λ is the ratio of the X and Y variances (Lord and Novick 1968).[14]

Under these assumptions, we can determine the equivalent size of a sample, i.e. the size that would yield the same variance of the sample mean if there were no measurement error: $n^* = \lambda^2 \cdot n$. If there were no error, equally precise estimates could be obtained with smaller samples (for instance, by 36 %, $(1 - \lambda^2)$, with a reliability index $\lambda = 0.8$).

In correlation analysis, if measurement error on X is assumed to be uncorrelated with X and with another variable Z, measured free of error, then the correlation coefficient between X and Z is attenuated with intensity proportional to the reliability index of Y: $\rho_{Y,Z} = \lambda_Y \rho_{X,Z}$. If Z is also measured with error, $W = Z + \eta$, with the η error of the same type as above and uncorrelated with ε, the correlation coefficient is attenuated even more: $\rho_{Y,W} = \lambda_Y \lambda_W \rho_{X,Z}$. In simple regression analysis too, measurement errors in independent variables lead to a downward bias in the parameter estimates (attenuation). In a multiple-regression context, measurement errors in independent variables still produce bias, but its direction can be either upward or downward. Random measurement error in the dependent variable does not bias the slope coefficients but does lead to larger standard errors.

The foregoing makes it clear that even unbiased and uncorrelated measurement errors may produce serious estimation problems.

How can we get a measure of the reliability of survey variables? A first possibility for time-invariant variables is the use of information collected over time on the same units (panel). In our survey half the sample is composed of panel households. If we assume that the measures of time invariant variables are independent (a plausible assumption for a survey conducted at two-year intervals), a comparison over time gives an indication of reliability.

[14] A reliability index evaluates the degree to which an instrument gives consistent results; "reliability" does not imply the accuracy of the measurement, i.e. its truthfulness. A reliable measurement device is not necessarily accurate, as for instance in case of correct and consistent recording of false information (Hand et al. 2001).

Let Y_s and Y_t be the values observed in two subsequent waves, with additive errors: $Y_s = X + \varepsilon_s$ and $Y_t = X + \varepsilon_t$. Under the assumptions that

$$E(\varepsilon_s, \varepsilon_t) = 0 \text{ and } E(X, \varepsilon_s) = E(X, \varepsilon_t) = 0, \quad \forall\, s, t = 1, \ldots, T, \quad s \neq t, \qquad (5)$$

the correlation coefficient between the two measurements Y_s and Y_t equals the square of the reliability index: $\rho_{Y_s,Y_t} = \lambda^2$. If there is no measurement error, the coefficient equals 1. Hence, a reduction in the precision of the data collection process or in the reliability of the respondents' answers lowers the correlation coefficient.

If we consider the surface area of the primary dwelling (computed only for households who did not move and did not incur extraordinary renovation expenses between the two survey waves), the correlation coefficient is 0.65 (and the reliability index $\lambda = 0.80$). For the year of house construction, the correlation coefficient is still lower ($\rho = 0.55$); in 73 % of the cases, the spread is less than five years, but sometimes it is much greater, probably reflecting response difficulties for houses that have been heavily renovated.

Another variable that is subject to inconsistency is the year when the respondents started working. The usual problems of recall are presumably aggravated in this instance by a certain degree of ambiguity in the question: it is not clear whether occasional jobs or training periods should be included or not. Out of 6,708 individuals who answered the question both in 2006 and 2008, 40.6 % gave answers that do not match; linear correlation was only 0.71.

All these examples underscore the great importance and the difficulty, in surveys, of framing questions to which respondents can provide reliable answers. It is not only a problem of knowledge and memory. There may also be a more general ambiguity in definitions (how to count a garden or terrace in the surface area of a house? Should the walls be included?), which can be limited (say, by instructing both interviewers and respondents) but cannot be eliminated.

Dealing with categorical variables complicates the study; in fact the models presented above are no longer adequate. An index of reliability for categorical variables can be constructed using two measures (Y_1 and Y_2) on the same set of n units. The fraction of units λ^* classified consistently is a reliability index (Biemer and Trewin 1997). Analytically, λ^* is given by

$$\lambda^* = \frac{tr(F)}{n} = \frac{\sum_{i=1}^{n} f_{ii}}{n}, \qquad (6)$$

where F is the cross-tabulation of Y_1 and Y_2 whose generic element is f_{ij} and $tr(.)$ is the trace operator, i.e. the sum of the diagonal elements.

However, the index λ^* does not take account of the fact that consistent answers could be partly random: if the two measures Y_1 and Y_2 are independent random variables, the expected share of consistent units is $\sum_{i=1}^{n} f_{i.} f_{.i}/n$. A reliability index that controls for this effect is Cohen's κ (Cohen 1960) that can be obtained by normalizing the share of observed matching cases with respect to the expected share, on the assumption that the two measurements of Y_1 and Y_2 were independent

Table 1 Reliability of type of high school degree, 2006–2008. Percentages

2008	A	B	C	D	E	F	Total
2006							
A. Vocational school	4.9	4.2	0.5	0.1	0.4	0.4	10.5
B. Technical school	4.1	44.1	2.4	0.3	0.6	1.1	52.7
C. Specialized high schools (*Licei*)	0.7	1.8	15.6	0.3	0.4	0.1	19.0
D. Art schools and institutes	0.1	0.1	0.2	2.0	0.2	0.0	2.6
E. Teacher training school	0.5	0.5	0.4	0.0	11.8	0.1	13.3
F. Other	0.3	0.5	0.4	0.0	0.2	0.4	1.9
Total	10.6	51.3	19.5	2.7	13.7	2.2	100.0
Reliability index λ^* (consistent answers)	88.7	84.2	92.7	98.7	96.6	96.7	78.8
Cohen's κ	40.1	68.4	76.5	74.8	85.4	17.8	68.0

$$\kappa = \frac{\lambda^* - \sum_{i=1}^{n} f_{i.} f_{.i}/n}{1 - \sum_{i=1}^{n} f_{i.} f_{.i}/n}. \tag{7}$$

Both λ^* and κ can also be applied to assess the reliability of all the categories of the qualitative variables, enabling us to pinpoint the main classification problems.[15]

If we compare the information on the type of high school diploma reported in the 2006 and 2008 waves, we find that about 20 % of the responses differ ($\lambda^* = 78.8$, Table 1). The transition matrix shows that a large part of the inconsistencies are between vocational and technical schools (4.1 and 4.2 %). In fact, the *Technical school* category reveals the lowest, but still high, reliability index $\lambda^*_B = 84.2$. However, once the correction for random consistent answers is considered the Cohen's measure of reliability turns out to be $\kappa = 68.0$. Moreover, the residual *Other* and *Vocational school* categories appear to be quite unreliable ($\kappa_F = 17.8$ and $\kappa_A = 40.1$).

Unfortunately, most of the SHIW variables vary over time, so their reliability cannot be measured by these techniques. More sophisticated instruments are required to distinguish actual changes from those induced by wrong measurements. A simple model allowing the estimation of the reliability index on time-varying quantities has been proposed by Heise (1969). The Author showed that, under mild conditions, real dynamics can be disentangled from measurement errors by taking three separate measurements of the economic variable on the same panel units.

Let X_1, X_2 and X_3 be the true unobservable values of the variable X during periods 1, 2, and 3, and Y_1, Y_2 and Y_3 be the corresponding observed measures. In order to apply the Heise method we assume that

$$Y_t = X_t + \varepsilon_t \quad \forall\, t = 1, 2, 3 \tag{8}$$

[15] Several indexes have been proposed for assessing the reliability of two or more measures (Krippendorff 2004). The so-called "weighted κ" has been proposed when the researcher may consider some disagreements less important than others (i.e. in the case of ordinal data). The Krippendorff's α is a more general index that can be applied on two or more repeated observations, on any metric (nominal, ordinal, interval, ratio), and on data with missing values (Krippendorff 2007).

and the dependency structure between X_1, X_2 and X_3 follows a first-order autoregressive model (not necessarily stationary) as

$$X_1 = \delta_1,\ X_2 = \beta_{2,1}X_1 + \delta_2, \ldots, X_3 = \beta_{3,2}X_2 + \delta_3 \tag{9}$$

where $\beta_{t,t-1}$ is the autoregressive coefficient and δ_t is a classical idiosyncratic error. We further impose that the innovation ε_t follows a white noise process and that the level of reliability of a given variable does not vary over time. On these assumptions the estimate of reliability can be derived from the following simple relation

$$\lambda^2 = \frac{\rho_{Y_1Y_2}\rho_{Y_2Y_3}}{\rho_{Y_1Y_3}}. \tag{10}$$

The intuition is that if measurement errors are independent over time and are not correlated with the underlying variable, then the absolute value of the estimated autocorrelation coefficients is lower than it would be if the observed value does not include measurement error. In fact, the method proposes an estimate of measurement reliability by comparing the product of one-step correlations $\rho_{Y_1Y_2}$ and $\rho_{Y_2Y_3}$ with the two-step correlation $\rho_{Y_1Y_3}$. Without measurement error, the product $\rho_{Y_1Y_2} \cdot \rho_{Y_2Y_3}$ would be equal to $\rho_{Y_1Y_3}$. As the intensity of measurement error is actually proportional to the square of $\rho_{Y_1Y_3}$, we can derive an indicator of measurement reliability by separating out the part that the model attributes to the actual variation of the underlying quantity.

In line with Biancotti et al. (2008), Table 2 reports the reliability indexes computed on three consecutive survey waves for the main variables, starting with 1989–1991–1993 and ending with 2004–2006–2008. The reliability estimate for income (on average 0.87) is higher than for net wealth and consumption (both averaging about 0.80).[16] Among the income components, higher index numbers are found for pension and transfer and for wage and salary (both around 0.95); incomes from self-employment or capital show lower values (around 0.80). As to the wealth components, greater reliability is found for real assets (on average 0.82), and in particular for primary residences (0.90), and lesser for financial assets (0.65).

These results are useful from three different perspectives. First, they allow the many researchers who use the survey to take this aspect properly into account, i.e. by selecting, among similar economic indicators, the most reliable. This benefit may also extend to other, similar surveys, which are likely to be affected by the same issues. Second, our results can help data producers for this kind of survey to find ways of reducing this kind of error; in fact, the difficulties discussed here are not specific to the SHIW data acquisition procedures. Quantifying their impact and determining their causes are essential preliminaries to improving survey procedures. Third, our conclusions can hopefully serve as standard practice for data producers and a blueprint for quality reporting.

[16] As noted, a reliability index does not measure the "closeness" of the reported to the true value, but only the variability of the measure. This implies that a systematic bias (for example due to consistent underreporting) will not be reflected in the Heise index.

Table 2 Heise reliability indexes of the main variables in the SHIW, 1989–2008

	1989 1991 1993	1991 1993 1995	1993 1995 1998	1995 1998 2000	1998 2000 2002	2000 2002 2004	2002 2004 2006	2004 2006 2008	Average
Net income	**0.89**	**0.94**	**0.89**	**0.84**	**0.89**	**0.85**	**0.80**	**0.82**	**0.87**
Wages and salaries	0.97	0.98	0.96	0.94	0.94	0.94	0.91	0.95	0.95
Pensions and transfers	0.97	0.99	0.96	0.93	0.93	0.94	0.97	0.90	0.95
Income from self-employment	0.89	0.97	0.84	0.71	0.75	0.81	0.73	0.82	0.82
Income from capital	0.82	0.79	0.81	0.77	0.79	0.78	0.74	0.81	0.79
Net wealth	**0.80**	**0.74**	**0.76**	**0.87**	**0.86**	**0.82**	**0.86**	**0.85**	**0.82**
Real assets	0.80	0.72	0.73	0.88	0.89	0.82	0.89	0.85	0.82
Financial assets	0.66	0.81	0.93	0.68	0.46	0.62	0.46	0.61	0.65
Financial liabilities	0.67	0.84	0.88	0.67	0.73	0.81	0.79	0.77	0.77
Consumption	**0.85**	**0.81**	**0.79**	**0.74**	**0.82**	**0.77**	**0.76**	**0.86**	**0.80**

3.3 Measurement Errors: Underreporting

In household surveys on income and wealth, the most significant type of measurement error is the voluntary underreporting of income and wealth. This type of error can produce severe bias in estimates, and special techniques are required to overcome this effect.

To evaluate the underreporting problem, a useful approach is to compare the survey estimates with other sources of data such as the National Accounts, administrative registers, fiscal data, and other surveys. For example, the number of dwellings declared in the survey differs significantly from the number owned by households according to the census.[17] On the basis of this evidence, underreporting by households could amount to as much as 20 or 25 % of all dwellings.

Further, underreporting is not constant by type of dwelling. While owner-occupied dwellings (principal residences) appear to be always declared, underreporting of other real estate owned proves to be very substantial. The SHIW itself allows a comparison between the estimate of the total number of houses owned by households and rented to others and the corresponding estimate drawn from the number of households living in rented dwellings.[18] In practice, the underestimation here appears to be very severe, as much as 60 or 70 %.

The estimates of real and financial wealth also appear to be underestimated by comparison with the aggregate accounts (Banca d'Italia 2010a). The bias is greater for financial assets, and underreporting is larger for less commonly held assets (equity and investment fund units). This suggests that unadjusted sample estimates are biased and that this distortion is not uniform across segments of the population.

[17] The number of dwellings owned by individuals is taken from the most recent census and updated using data from CRESME (CRESME 2010) on new buildings (owned by natural persons).

[18] Note that owing to sampling variance, even without underreporting the two estimates, though close, should not be exactly the same (Cannari and D'Alessio 1990).

How can we learn more about this, and how can we adjust the estimates accordingly? One way of assessing the credibility of the survey responses is to ask for the interviewers' own impression. That is, in the course of the interviews they are requested to look out for additional information, making a practical comparison between the household's answers and the objective evidence they can see for themselves: type of neighbourhood and dwelling, the standard of living implied by the quality of furnishings, and so on.

In the 2008 survey, credibility is satisfactory overall (an average score of 7.6 out of 10) but not completely uniform. The highest scores are for the better educated and for payroll employees (7.9 and 7.8, respectively), the lowest for the elderly and the self-employed (7.4 and 7.3, respectively).

The correlation coefficient between the credibility score and the declared values of income, financial assets and financial liabilities is positive and significant, but small. The use of this type of information is of little help for the adjustment of the estimates. For example, considering only the sub-sample of households with credibility better than 5 (around 90 % of the sample), average household income rises by just 1.1 %. The adjustment is a bit larger (2.8 %) considering only the households that score 7 or more. In these two cases, the wealth adjustments are respectively 0.8 and 3.2 %; the adjustment for financial assets is greater (between 4 and 11 %).

Taking a completely different approach, underreporting can be analysed by statistical matching procedures. Cannari et al. (1990) performed statistical matching between the SHIW answers and the data acquired by means of a specific survey conducted by a commercial bank on its customers. Under the hypothesis that the bank clients report the full amount of financial assets held, as customers are likely to trust their bank, the Authors estimated the amount of financial assets held by the households in the SHIW database.[19] The study concluded that the survey respondents tend to underreport their assets quite significantly. The underreporting involved several different components. Some households, in fact, do not declare any bank or postal accounts, and hence the ownership of financial assets is underestimated. This behaviour was determined to result in an underestimation of about 5 %; it was more frequent among the poorer and less educated respondents. Underestimation due to non-reporting of single assets, i.e. the omission of assets actually held, involved a further 10 % of assets. But the bulk of the underreporting concerned the amounts of the assets declared. The study found that for a declared value of 100, households actually held assets worth 170.

Applying this correction, the total amount of financial assets owned by households doubled. The discrepancy with respect to the financial accounts was sharply reduced, but a significant gap remained, presumably deriving from definitional differences and the very substantial asset holdings of the tiny group of very wealthy

[19] On the assumption that the probability of declaring an asset not actually held is zero, the conditional probability of not declaring an asset held is simply obtained by using marginal probability: $p_{h/nd} = 1 - (1 - p_h)/(1 - p_d)$. The marginal probabilities can be estimated on the two samples separately.

households, which are not properly represented in sample surveys. The adjustment ratio for financial assets, finally, was higher among the elderly and the self-employed.

Another matching experiment, based on the same data but with different methods (Cannari and D'Alessio 1993), confirmed the foregoing results. The experiment also showed that the Gini concentration index of household wealth was not seriously affected by the adjustment procedures (from 0.644 to 0.635 for 1991).

In a recent paper on this topic, D'Aurizio et al. (2006) use an alternative method and data drawn from a different commercial bank. On average, the adjusted estimates are more than twice the unadjusted data and equal to 85 % of the financial accounts figures. The adjustments are greatest for the households whose head is less educated or retired.

Neri and Zizza (2010) propose different approaches to correct for underreporting of household income. To adjust the estimates for self-employed households, the procedure uses the ratio of the value of the primary residence to labour income; this approach is a variant of the one proposed by Pissarides and Weber (1989), based on the ratio of food expenditure to income. The ratio of the value of homes to labour income is estimated first for public employees, whose answers are presumed not to be underreported. The estimated parameters are then applied to the self-employed (the value of houses is assumed to be reported correctly by both types of respondent). On this basis the estimated average income from self-employment is 36 % greater than the unadjusted figure. To adjust income from financial assets, the authors used the (D'Aurizio et al. 2006) methodology for the correction of financial stocks, simply applying a return rate to the adjusted capital stock. It was found that on average this adjustment tripled the reported income. The increase in liabilities was modest (just 9 %). As to the income from real estate, they used the procedure developed by Cannari and D'Alessio (1990), which adjusts the number of declared second homes to the Census. The income from actual and imputed rents increased on average by 23 %. Income sources from other labour activities was adjusted on the basis of the Italian part of the European Union Statistics on Income and Living Conditions (EU-SILC), which includes information from administrative and fiscal sources. With this adjustment, additional payroll and self-employment income increased by 3 and 4 % points respectively. Overall, the adjustment procedures produce an estimate of total family income about 12 % greater than the declared value (between 2 and 4 times the corresponding sampling errors). In summary, analysis of the discrepancy between the survey figures and the financial accounts shows the simultaneous presence of non-response, non-reporting and underreporting. The underestimation of financial assets and liabilities due to non-participation in the survey appears to be less substantial than that caused by non-reporting and underreporting.

In the 2010 survey, the SHIW tried the *unmatched count technique* (Raghavarao and Federer 1979) for eliciting honest answers on usury, a serious problem mainly for small businesses and poor households but a phenomenon on which no reliable information is available. The technique uses anonymity to get a larger number of true answers to sensitive or embarrassing questions. In this case, the respondents are randomly split into two groups, A and B. The control group B is asked to answer a set of k harmless binary questions $X_1, \ldots, X_k$, while the treatment group A has one additional question Y (the sensitive one). The respondents in both groups are to reveal only the number of applicable activities or behaviors, not to respond specifically to each item. Hence, the answers have the forms of $S_B = X_1 + X_2 + \cdots + X_k$ and $S_A = S_B + Y$ for respondents belonging to A and B group respectively. With the unmatched count, the number of people who answered "yes" to the sensitive question is estimated by comparing the two mean values: $\overline{Y} = \overline{S}_A - \overline{S}_B$. Under certain conditions, researchers can also perform regressions on this type of data.[20]

4 Concluding Remarks

This work has described the research done at the Bank of Italy on non-sampling errors in the SHIW to bring out the most common problems in household surveys. These errors are frequent and constitute the largest part of the total error. We gauge the impact of non-participation in the survey, classic measurement error and under-reporting, and describe some practical procedures for correcting these error sources. We show that the correction procedures often depend on the specific assumptions. For this reason the techniques are more in the nature of tools that a researcher can legitimately use than of standard practices for the production of descriptive statistics, such as those reported in the official Bank of Italy reports.

As survey designers, we have shown that it is simply essential to collect additional information, beyond that strictly related to the content of the survey. In the SHIW, we acquire information on: the households not interviewed; the effort needed to acquire the interviews; the time spent on the interviews; the credibility of answers; and the characteristics of the interviewers themselves. All these data can help us to grasp the extent and the causes of the various types of non-sampling error.

The analysis may serve to suggest more effective survey design. In fact, we have shown the lower response rate observed for wealthier households, which the usually employed stratification and post-stratification criteria are not able to correct properly. The availability of data on the average market value of houses by neighbourhood within the main cities suggests that serious consideration should be given to revising these criteria. Another solution might be the over-sampling of wealthier households to improve the efficiency of some overall estimators.

Specific techniques for collecting sensitive information are available. More generally, the questionnaire should be designed to include careful evaluation of various aspects of apparently less problematic questions as well.

Another matter for further research, on which work is under way, is interviewer effects: heterogeneous performances among interviewers in terms of response rate and measurement error. The results could help us to improve selection and training procedures.

[20] Another technique for this purpose is the *randomized response technique* proposed by Warner (1965). However, this procedure is too cumbersome for a multi-purpose survey like the SHIW.

The work also showed that the sample estimates for income and wealth are seriously affected by underreporting, in spite of the efforts to overcome respondents' distrust. This evidence suggested increasing the share of panel households, which was accordingly raised from 25 % in 1991 to 55 % in 2008. Panel households, in fact, are better motivated to give truthful responses. The average credibility score for the panel households is greater than for households interviewed for the first time (7.73 as against 7.44 in 2008). However, while it may improve response credibility, increasing the panel proportion may reduce the coverage of particular population segments (e.g. young households) and worsen sample selection due to unit non-response. The terms of this trade-off need to be carefully evaluated.

As survey data users, we are aware that knowledge of the types of non-sampling errors can greatly improve both the specification of the empirical model and the interpretation of the results. In conclusion, we urge that in using surveys data practitioners maintain a critical reserve concerning the possible non-sampling errors affecting this type of data.

Acknowledgments We would like to thank the three anonymous referees for their valuable comments and suggestions to improve the quality of the paper. The views expressed are those of the authors and do not necessarily reflect those of the Bank of Italy.

References

Banca d'Italia. (2010). Household Wealth in Italy in 2009. *Supplements to the Statistical Bulletin* (new series) (Vol. 67), Banca d'Italia.

Banca d'Italia. (2010). Italian household budgets in 2008. *Supplements to the Statistical Bulletin* (new series) (Vol. 8), Banca d'Italia.

Biancotti, C., D'Alessio, G., & Neri, A. (2008). Measurement error in the bank of italy's survey of household income and wealth. *Review of Income and Wealth, 54*, 466–493.

Biemer, P.P., & Lyberg, L. (2003). *Introduction to Survey Quality*. Wiley Series in Survey Methodology. Hoboken: Wiley.

Biemer, P. P., & Trewin, D. (1997) A review of measurement error effects on the analysis of survey data. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey Measurement and Process Quality* (pp. 603–632). New York: Wiley-Interscience.

Blom, A. (2011). Measuring interviewer effects across countries and surveys (pp. 18–22). Paper presented at the *Fourth conference of the european survey research association*, Lausanne.

Cannari, L., & D'Alessio, G. (1990). Housing assets in the bank of Italy's survey of household income and wealth. In C. Dagum, M. Zenga, (Eds.), *Proceedings of Income and Wealth Distribution, Inequality and Poverty* (pp. 326–334). Berlin: Springer.

Cannari, L., & D'Alessio, G. (1993). Non reporting and under reporting behaviour in the bank of Italy's survey of household income and wealth. In *Bulletin of the International Statistics Institute 49th Session* (pp. 395–412). Firenze: International Statistical Institute.

Cannari, L., D'Alessio, G., Raimondi, G., & Rinaldi, A. (1990). Le attività finanziarie delle famiglie italiane. *Temi di Discussione* (Vol. 136), Banca d'Italia.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

CRESME. (2010). Il Mercato delle costruzioni al 2011. *Rapporto congiunturale e previsionale* (Vol. 18), CRESME, Roma.

D'Alessio, G., & Faiella, I. (2002) Nonresponse behaviour in the bank of Italy's survey of household income and wealth. *Temi di Discussione* (Vol. 462), Banca d'Italia.

D'Aurizio, L., Faiella, I., Iezzi, S., & Neri, A. (2006). L'underreporting della ricchezza finanziaria nell'indagine sui bilanci delle famiglie. *Temi di Discussione* (Vol. 610), Banca d'Italia.

De Luca, G., Peracchi, F. (2011). Estimating engel curves under unit and item nonresponse. *Journal of Applied Econometrics*. doi:10.1002/jae.1232

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. Cambridge: MIT Press.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica, 47*, 153–161.

Heise, D. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review, 34*, 93–101.

Iannacchione, V. G. (2003). Sequential weight adjustments for location and cooperation propensity for the 1995 national survey of family growth. *Journal of Official Statistics, 19*, 31–43.

Kalton, G., & Flores Cervantes, I. (2003) Weighting methods. *Journal of Official Statistics, 19*, 81–97.

Krippendorff, K. (2004). Reliability in content analysis: some common misconceptions and recommendations. *Human Communication Research, 30*, 411–433.

Krippendorff, K. (2007). *Computing krippendorff's alpha reliability*. Departmental papers 43, Annenberg School for Communication, University of Pennsylvania.

Lessler, J., & Kalsbeek, W. (1992). *Nonsampling Error in Survey*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Little, R., & Rubin, D. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R., & Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology, 31*, 161–168 (2005).

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Neri, A., & Zizza, R. (2010). Income reporting behaviour in sample surveys. *Temi di Discussione* (Vol. 777), Banca d'Italia (2010).

Pissarides, C.A., & Weber, G. (1989). An expenditure-based estimate of Britain's black economy. *Journal of Public Economics, 39*, 17–32 (1989).

Raghavarao, D., & Federer, W. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society Series B, 41*, 40–45.

Särndal, C. E., & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley series in survey methodology, Chichester: Wiley.

Singh, A. C., & Mohl, C. A. (1996). Undertsanding calibration estimators in survey sampling. *Survey Methodology, 22*, 107–115.

Warner, S. L. (1965). Randomized-response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*, 63–69.