MEASUREMENT ERROR IN THE BANK OF ITALY'S SURVEY OF HOUSEHOLD INCOME AND WEALTH

Claudia Biancotti, Giovanni D'Alessio and Andrea Neri
(Bank of Italy, Research Department)*

This paper is aimed at evaluating the incidence of measurement error on the main variables collected in the Bank of Italy's Survey of Household Income and Wealth (SHIW). The results are especially relevant to researchers using the data for economic analysis, since they need to take data quality into account. Moreover, a thorough knowledge of the problems affecting the survey gives indications for improvements in its design and implementation.

Where time-invariant variables are concerned, measurement error is studied by assessing the degree of inconsistency of answers given by panel households in subsequent survey waves. In the case of quantities that have an actual variation in time, such as income or wealth, the Heise (1969) model is applied; if data from at least three waves are available, we can separate the true dynamics from the noise of measurement error, under assumptions that are fairly mild. The essay also touches upon the role of fieldwork conditions, interviewer and respondent features in the determination of data quality.

*We would like to thank Luigi Cannari, Ivan Faiella, Grazia Marchese, Luigi Federico Signorini and two anonymous referees for their useful insights. The opinions expressed in this paper are those of the authors and should not be attributed to the Bank of Italy.

1. Introduction

Estimates that are based on sample surveys are subject to a number of possible errors. A first source of inaccuracy is implicit in the nature of the inferential process that yields population parameters on the basis of a set of sampled units; a type of disturbance known as *sampling error* arises, whose incidence can be evaluated precisely if we are aware of some features of the sample (e. g. size, design) and of the population (e. g. distribution of the variable we are interested in).

Other causes of imprecision involve the process of measurement and estimation; the resulting mistakes are known as *non-sampling errors*, the sometimes unavoidable costs of transforming a theoretical scheme into an actual survey.

Broadly speaking, the literature on measurement error focuses on the problems relating to the following aspects:

- sample composition, as a consequence of incomplete sampling frames (non-coverage) or failure to participate in the survey on the part of some sampled units (non-response);
- discrepancies between recorded data and "true" data, originating from response error or oversights in the processing phase prior to estimation.

The effects induced by some types of error on the estimated values of aggregates in the Bank of Italy's Survey of Household Income and Wealth (SHIW from here on) have been studied in the past. For example, sampling errors are normally published along with survey results (Bank of Italy, 2002); the consequences of non-response on the most important estimates have been assessed[1]; efforts have been made to evaluate the magnitude of under-reporting of assets and income[2].

---

[1] Cannari and D'Alessio (1992) analyze the behavior of panel households and find that non-response is a common trait in large cities and in Northern Italy. The participation rate decreases with income, and increases with household size. D'Alessio and Faiella (2002) confirm that well-off households and those headed by educated individuals are harder to interview, while households residing in Central Italy and headed by an individual in the central age groups are more likely to participate in the survey.

[2] The value of housing, which accounts for most of real wealth, appears to be underestimated by 20%; the figure is higher when referred to non-primary (vacation etc.) housing only. Financial assets are also seriously exposed to under-reporting (Cannari and D'Alessio, 1993; Cannari, D'Alessio, Raimondi and Rinaldi, 1990). Income deriving from self-employment and from capital also appears to be severely underestimated (Cannari and Violi, 1995; Brandolini, 1999).

These analyses notwithstanding, several areas of the data quality territory remain relatively uncharted, especially in relation to response error. In a survey of income and wealth, under-reporting remains the most dangerous cause of differences between reported values and “true” values, because it can heavily condition the estimates of means. However, the literature also cites several other factors that potentially reduce data quality.

The questionnaire is not a neutral instrument: the order and wording of questions and the available response options influence the answers, especially (but not only) where opinions, expectations and other subjective items are concerned.

Interviewer behavior is also very important: there are a number of ways of asking the same question in a face-to-face setting, and each can induce a different psychological reaction, ultimately affecting the answer.

Further problems can arise from the respondent’s cognitive processes: hypothetical questions require some abstract reasoning, retrospective ones need an effort to recall events of the past[3]. Moreover, people may not actually know the exact answer to the questions they are asked, especially in cases (such as the SHIW) where response by proxy is allowed.

Following Groves and Couper (1998), general aspects such as motivation of the respondent and willingness to give time and effort for a survey should also be assumed to influence data quality.

Finally, the use of a Computer-Assisted Personal Interviewing electronic interface (CAPI) – complete with range controls, consistency assessment, and outlier detection tools – instead of a printed form can influence the answers.

The possible causes of response error, as summarized above, appear to be too numerous to be tackled in a single essay. We will therefore concentrate on the impact that certain features of fieldwork operations, of the interviewers and of the respondents have on data quality.

The statistical analyses that follow are based entirely on survey data; hence, they may not meet the strict randomization criteria needed for controlled experiments. Some caveats

---

[3] For example, a hypotetical question entailing a certain effort on the part of the respondent is asked to homeowners in the SHIW (Bank of Italy, 2002): *Assuming you wanted to rent this dwelling, what monthly rent do you think could be charged? Do not include condominium charges, heating or other sundry expenses*. Memory problems could arise in questions such as this one, directed to pensioners: *Recall when you began to receive your pension. What percentage of your last wage payment (monthly average earnings, if self-employed) was your first pension payment?*

apply to the conclusions: they are only as reliable as the models used to eliminate possible sources of noise. This will be discussed later on.

As a further warning, note that this paper is a first exploration of a subset of measurement error issues in the SHIW: it gives some elements to evaluate the magnitude of imprecisions in the data collection process, and on the possible reasons why they exist. It does not assess their consequences on the estimation of mean values, regression coefficients and other statistics[4].

The paper is structured as follows. Section 2 briefly describes the SHIW, with a special focus on what is relevant for data quality. Section 3 proposes a methodology for evaluating the degree of reliability of collected data. Section 4 presents some descriptive statistics on measurement error for the main SHIW aggregates (income, wealth, consumption) and their individual components. Section 5 puts forward models that try to explain the inconsistence in answers provided over the years by panel households. Section 6 concludes.

2. The Bank of Italy's Survey of Household Income and Wealth

Since 1962, the Bank of Italy conducted a survey on household budgets, examining economic behavior at the micro level.

In the recent waves, the sample size has been of about 8,000 households. The design is two-stage; 300 municipalities, stratified by region and population, are selected at random, and then households are drawn from the municipal registers.

Starting from the 1989 wave, a part of the sample (now roughly 50 per cent) is made up of households with previous experience of SHIW participation, the so-called panel households. This structure permits the study of dynamic phenomena such as income, wealth and employment mobility.

---

[4] A vast literature exists on the effects of measurement error on survey-based estimates. Even a completely random error, although devoid of consequences on the estimation on mean values and population totals, and affecting variance in a way that can be corrected by modifying the sample size, distorts a number of statistics (e. g. quantiles or linear regression coefficients), in ways that have to be studied by way of non-trivial models such as the ones presented by Biemer and Trewin (1997). Quite often the error is not completely random: we will see in subsequent paragraphs that it can depend on fieldwork, interviewer and/or respondent features. In this case, complex models are necessary to predict the effect of said error on the estimate; a separate paper should be devoted to this problem and possible remedies.

The questionnaire always addresses the following topics: demographic structure of the household, educational and occupational features of each member, individual income, household wealth and consumption, housing. Variable monographic sections are added on the basis of specific needs.

The survey is materially implemented by a specialized company hired by the Bank of Italy. Those who agree to participate are interviewed personally in their homes, often with the CAPI interface.

Most aspects of fieldwork are documented: information is collected on interviewer features (Table 1), on CAPI use (Table 2), on the presence of household members other than the head[5] during the interview (Table 3), on the date and time of the interview and its length (Figure 1).

The typical interviewer is female, a professional (interviewing is her main job), slightly over 40, with a high school degree. In the Northern regions, which are the richest, interviewer turn-over is higher: the mean value of experience in the job is 9 years, while in the South it is 11.5 years. These figures, together with the higher incidence of interviewers who are not professionals (36.9% against 25.3%) and who hold junior high school degrees only (16.5% against 6.1%), seem to reflect the difficulty of finding work encountered by high school leavers in the South. On average, 67% of interviews are carried out with the CAPI method, which is largely present in the areas where most interviewers are professionals. On one hand, the company in charge of the survey might be more inclined to endow with computers stable employees than short-term ones; on the other, people who do not interview full-time might not want to put effort in learning how CAPI is used.

[5] The head of the household is defined as the person responsible for the household's economic decisions.

Table 1

INTERVIEWER FEATURES, 2000

(percentages)

	North	Center	South	Total
Sex				
Male	25.2	15.0	24.2	23.1
Female	74.8	85.0	75.8	76.9
Educational qualification				
Junior high school	16.5	7.5	6.1	10.7
High school	69.9	72.5	83.8	76.0
University or more	13.6	20.0	10.1	13.2
Age class				
Up to 30	12.6	7.5	17.2	13.6
31 to 40	25.2	22.5	33.3	28.1
41 to 50	23.3	32.5	30.3	27.7
51 to 65	32.0	32.5	19.2	26.9
Over 65	6.8	5.0	0.0	3.7
Professional interviewer				
No	36.9	37.5	25.3	32.2
Yes	63.1	62.5	74.7	67.8
Total	100.0	100.0	100.0	100.0
Experience (years)	9.0	10.6	11.5	10.3
Number of SHIW waves carried out	1.8	1.7	2.1	1.9

Table 2

USE OF THE ELECTRONIC QUESTIONNAIRE, 2000

(percentages)

Area	Paper	CAPI	Total
North	47.8	52.2	100.0
Center	20.3	79.7	100.0
South	21.8	78.2	100.0
Italy	33.0	67.0	100.0

Most heads of household are interviewed personally[6]; so are 27.9% of their spouses or live-in partners, while the rest of the members are normally not present during the interview; the rate of personal response decreases with the number indicating the position in the household, which is declared by the household head during the interview.

[6] The presence of the head of household is a necessary condition for the interview. The few cases of absence correspond to exceptional situations, such as the death of the household head between the end of the reference year and the day of the interview. The presence of the other members is recorded only on income-related annexes to the main questionnaire; the information is therefore unavailable for non-earners, 39.4% of the sample. As a consequence, some of the estimates in Table 3 are biased downwards.

Table 3

PERSONAL AND PROXY INTERVIEWS BY HOUSEHOLD MEMBER, 2000

(percentages)

Position in the household	Interviewed personally	Interviewed by proxy	Unknown	Total
1	99.8	0.2	0.0	100.0
2	27.9	35.0	37.1	100.0
3	5.9	29.0	65.1	100.0
4	3.0	16.9	80.1	100.0
5	2.9	15.6	81.5	100.0
6	2.4	18.2	79.4	100.0
7	2.4	19.5	78.0	100.0
8	6.7	26.7	66.7	100.0
9	0.0	0.0	100.0	100.0
Total	42.1	18.5	39.4	100.0

On average, an interview takes 55 minutes, with a standard deviation of 19 minutes. Interview length is explained by socio-demographic features, such as the number of members and of earners (single-person households take only 46 minutes, five-member households take an hour, large families over 70 minutes), and by income levels (46 minutes for households earning less than 10,000 euros per year, 64 minutes for those over 40,000 euros)[7]. Operational choices also influence the amount of time needed to complete the interview: paper questionnaires take 58 minutes on average, the CAPI method 54.

The number of responding households per interviewer is distributed as shown in Figure 2 (non-parametric density estimate performed with an Epanechnikov kernel and Silverman (1986) optimal bandwidth). A certain variability emerges: the distribution mean is 33, and 75 per cent of the cases fall between 8 and 60. The asymmetry is justified by the fact that during the last weeks of fieldwork the best interviewers get "recovery" assignments in order to boost the response rate.

[7] The questionnaire is structured in such a way that each household member, each job held, each real estate asset requires a separate form.

Figure 1

DISTRIBUTION OF INTERVIEWS BY LENGTH, 2000
(minutes)

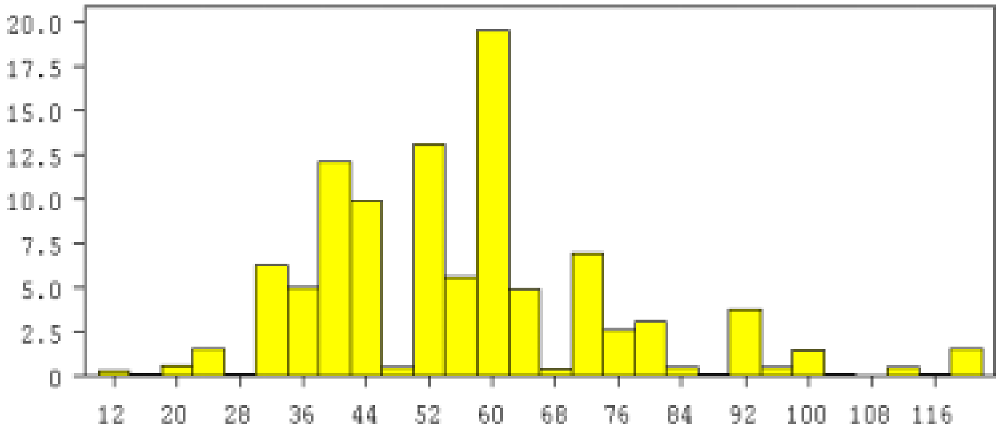
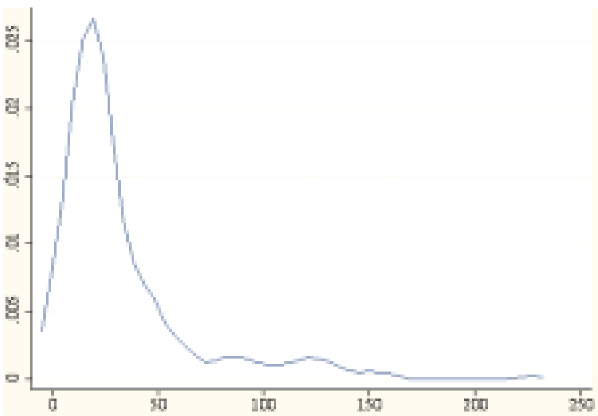


Figure 2

DISTRIBUTION OF INTERVIEWS PER INTERVIEWER, 2000
(number of interviews)



3. The evaluation of data reliability: methodology

Let the variable X be measured with error:

$$Y = X + e \quad (1)$$

The measure Y differs from the true value X by a random component with the following properties:

$E(e) = 0$ (2)

$E(X, e) = \sigma_{X,e} = 0$ (3)

$E(e, e) = \sigma^2_e$ (4)

This type of disturbance is called *homoskedastic, uncorrelated error*. Under these assumptions, the variance of Y is a biased estimator of the variance of X, since:

$\sigma^2_Y = \sigma^2_X + \sigma^2_e = \sigma^2_X / \lambda^2$ (5)

where

$$\lambda = \frac{\sigma_X}{\sigma_Y} \quad (6)$$

The coefficient λ is known as the *reliability index*[8]; it expresses the share of variability in Y that originates from the true phenomenon X (Lord and Novick, 1968)[9].

This index can be interpreted in several ways, taking into account the impact that measurement errors as described in (1)-(3) can have on different statistics. For example, the expected value of the measurement Y is an unbiased estimator of the mean of X: E(Y)=E(X)+E(e)=E(X). Still, the presence of an error induces a higher estimator variance:

$\sigma^2_{E(Y)} = \sigma^2_{E(X)} + \sigma^2_e/n = \sigma^2_{E(X)} / \lambda^2$ (7)

From a sampling point of view, (7) implies that λ allows us to determine the "effective" sample size $n^* = \lambda^2 n$, i. e. the size that would yield the same variance of the sample mean if there were no measurement error[10].

---

[8] Following, among others, Hand, Mannila and Smyth (2001), "A precise measurement procedure is one that has small variability […] [.A]n accurate measurement procedure, in contrast, not only possesses small variability, but also yields results close to what we think of as the true value. […] The reliability of a measurement procedure is the same as its precision. The former term is typically used in the social sciences whereas the latter is used in the physical sciences". A reliability index evaluates the degree to which an instrument, in our case the SHIW questionnaire, yields results that portray reality consistently; it does not indicate the instrument's truthfulness. We want to see what additional distance between data collected in different waves is introduced by measurement error, possibly net of actual changes in the quantities studied; a reliability index does not assess the distance between *collected* data and *true* data. Moreover, a precise measurement is not necessarily accurate, as shown by the case of correct and consistent recording of false information; reliability indexes are not able to spot the presence of phenomena such as systematic under-reporting.

[9] In the rest of the paper we will use the reliability index as a descriptive parameter for the specific sample we are dealing with, not as an estimate of the corresponding population parameter. We only consider the variability introduced by measurement error, ignoring the variability connected to the sampling process.

[10] We might also say, equivalently, that $(1-\lambda^2)/\lambda^2$ is the additional cost introduced by measurement error; for example, a reliability index λ of 0.8 implies a rise in survey costs of 56%; if there were no error, estimates with the same precision could be obtained with a sample smaller by 36% $(1-\lambda^2)$.

Turning to correlation analysis, we can say that if measurement error on X is assumed to be uncorrelated with X and with another variable Z, measured free of error, then $\rho_{Y,Z} = \lambda_Y \rho_{X,Z}$. The correlation coefficient between X and Z is attenuated with intensity proportional to the reliability index of Y, the measure of X. If we had a measure W of the variable Z, affected by measurement error uncorrelated with Z, X and the error component on X, then we would have $\rho_{Y,W} = \lambda_Y \lambda_W \rho_{X,Z}$. Under these conditions, the coefficient of univariate linear regressions of the type $Z = \alpha + \beta_X X$ estimated on the basis of the observed variables Y and W are related to the true coefficients by way of $\beta_Y = \lambda_Y{}^2 \beta_X$.

We can estimate λ if we have two measurements of the same phenomenon[11]. Let Y_1 and Y_2 be such measurements, with additive errors[12]:

$Y_1 = X + e_1 \qquad Y_2 = X + e_2$ (8)

Let assumptions (2), (3) and (4) hold, supplemented by

$E(e_1, e_2) = \sigma_{e1,e2} = 0 \qquad E(X_t, e_{t'}) = \sigma_{Xt,et'} = 0 \ \forall t,t'$ (9)

Under these conditions, the correlation coefficient between the two measurements Y_1 and Y_2 equals the square of the reliability index:

$$\rho_{y1,y2} = \sigma_{y1,y2} / \sigma_{y1}\sigma_{y2} = \sigma^2_x / (\sigma^2_x + \sigma^2_e) = \sigma^2_x / \sigma^2_y = \lambda^2_y \qquad (10)$$

Panel households are interviewed every 2 years and sometimes they are asked questions they have answered in preceding waves. For these variables, if they are time-invariant, a quantification of measurement error can be obtained by applying (10) to the two answers.

If we are dealing with categorical variables, the estimation of (6) is not so straightforward, since the model in (1) is no longer adequate; it is not possible to represent

---

[11] What was stated in relation to regression coefficients implies that we may also estimate the reliability of a variable based on a single wave; i. e. under the conditions necessary for the use of instrumental variables (IV), and in the presence of univariate relations that are sufficiently significant and exhaustive, the reliability index can be calculated as the ratio between the OLS regression coefficient and the IV regression coefficient.

[12] The results presented from here on can be extended in a straightforward way to the case of error variance proportional to the variance of the observed quantity. This phenomenon can be described by a model that is similar to the purely additive one used above: Y = Xe, with e=1+u and E(u)=0 (heteroskedastic errors). If the *u* term, which expresses the “error share” of the variable measured, is still uncorrelated with X, then all the reasoning in this section can be replicated.

the error term as additive, and the hypotheses on the mean, variance and correlation of errors need to be revised.

Let X be a binary (dummy) variable with parameter p, and Y its measurement; and let θ and ϕ be the probabilities of measurement-related misclassification for the two categories X=1 and X=0: $\theta = pr(Y=0|X=1)$ and $\phi = pr(Y=1|X=0)$. It is easy to show that the estimate of p based on Y is biased, as in:

$$E(Y) = (1 - \theta)\, p + \phi\ (1 - p) \neq p \qquad (11)$$

unless the errors in the two categories exactly compensate one another, with $\phi = \theta\, p / (1\text{-}p)$.

It is interesting to note that the estimate of p is generally distorted when the two misclassification rates θ and ϕ coincide, which is one of the possible descriptions of measurement error. We have

$$E(Y) = (1 - \theta)\, p + \theta\ (1 - p) = p + \theta\, (1\text{-}2p) \qquad (12)$$

From (12) we can say that in the case of equal misclassification rates $\theta = \phi$ the expected value of Y estimates the mean of X without bias only if p=0.5. In all other cases, the bias $\theta(1\text{-}2p)$ pushes the estimate towards 0.5 as the error rate increases; the bracketed quantity is positive if p<0.5 and negative if p>0.5. Anyway, we can never be sure that $\sigma^2_x < \sigma^2_y$: it follows that definition (6) is not applicable to binary variables.

An alternative measure of reliability for categorical features measured twice on the same set is the fraction f of units that are classified consistently: λ^*=f. The index, however, does not take into account the fact that the frequency of inconsistent classifications always tends to go up with the variance of Y, $\sigma^2_y = p_y\,(1\text{-}p_y)$. This can be seen clearly if we consider two independent random variables Y_1 and Y_2 with the same mean, p_y. We expect to find $2p_y\,(1\text{-}p_y) = 2\sigma^2_y$ cases of non-matching values and $p^2_y + (1\text{-}p_y)^2 = 1\text{-}2\sigma^2_y$ cases of matching values. A version of the reliability index which controls for this effect (see Biemer and Trewin, 1997) can be obtained by normalizing the share of observed matching cases with respect to their expected incidence if the two measurements of Y were independent:

$$\lambda^{**} = (f - 2\sigma^2_y) / (1\text{-}2\sigma^2_y) \qquad (13)$$

It is useful to note that this index, like that presented in (6), gives no information about the possible bias of the estimator.

The indexes discussed so far allow us to derive a measure of response errors on variables that are measured twice and independently. In the SHIW context, this is the case of some phenomena that do not vary with time; since there is a two-year interval between interviews, the risk of contamination between different waves is very low (respondents most probably do not remember what they said)[13].

The analysis of measurement errors on the large majority of collected variables, especially the most interesting ones such as income and wealth, requires more sophisticated instruments. If a quantity varies with time, it is necessary to distinguish actual change from movements induced by wrong measurement.

The reliability of data on time-varying quantities can be assessed with the Heise (1969) method: provided we have at least three separate measurements of a variable on the same panel units (e. g. answers to the same question in three survey waves), under mild regularity conditions we are able to separate real dynamics from measurement error.

Let X_1, X_2 and X_3 be the true values of the variable X during periods 1, 2 and 3; Y_1, Y_2 and Y_3 are the corresponding measurements, for which the following equation applies:

$$Y_t = X_t + e_t \; \forall t \qquad (14)$$

In addition to this, let X_1, X_2 and X_3 be pairwise related through independent, first-order autoregressive models, which do not need to be stationary:

$$X_1 = \delta_1 \qquad (15)$$

$$X_2 = \beta_{21} X_1 + \delta_2 \qquad (16)$$

$$X_3 = \beta_{32} X_2 + \delta_3 \qquad (17)$$

where $\beta_{t+1,t}$ is the autoregressive coefficient and δ_t is the process innovation. Innovations are uncorrelated pairwise.

Assuming that the level of reliability of a given variable does not vary with time, the correlation coefficient between the observed values Y_t and Y_{t+1} can be written as:

$$\rho_{Yt,Yt+1} = \lambda_{Yt}\lambda_{Yt+1}\rho_{Xt,Xt+1} = \lambda_Y^{\;2}\rho_{Xt,Xt+1} \qquad (18)$$

The ratio between the coefficient observed and the one that would be observed in the absence of measurement error is therefore always smaller than 1 and equal to λ_Y^2:

$$\frac{\rho_{Yt,Yt+1}}{\rho_{Xt,Xt+1}} = \lambda_Y^{\;2} \qquad (19)$$

---

[13] Respondents probably remember that they participated in the survey two years before, and this can have an influence on their motivation or on their attitude towards giving information perceived as sensitive. In turn, this impacts on some types of error.

Since the true values are related by way of independent, first-order autoregressive processes we can say that for each t the following holds:

$$\frac{\rho_{Xt-1,Xt}\,\rho_{Xt,Xt+1}}{\rho_{Xt-1,Xt+1}} = 1 \qquad (20)$$

Substituting (19) into (20), the estimate of reliability can be written as:

$$\lambda_Y = \sqrt{\frac{\rho_{Yt-1,Yt}\,\rho_{Yt,Yt+1}}{\rho_{Yt-1,Yt+1}}} \qquad (21)$$

and more generally:

$$\lambda_Y = \sqrt[n]{\frac{\prod_{s=t}^{t+n} \rho_{Ys,Ys+1}}{\rho_{Yt,Yt+n}}} \qquad (22)$$

The main idea of the method is: if measurement errors are independent of time and of the underlying variable, the absolute value of the estimated autocorrelation coefficients turns out to be lower than what we would get if the observed value did not include measurement error. Assuming that the true values in the three periods X_1, X_2 e X_3 are related via first-order autoregressive models, the method proposed an estimate of measurement reliability by comparing the product of one-step correlations $\rho_{Y1,Y2}\,\rho_{Y2,Y3}$ with the two-step correlation $\rho_{Y1,Y3}$. If no measurement error existed, the quantity $\rho_{Y1,Y2}\,\rho_{Y2,Y3}$ would be equal to $\rho_{X1,X3}$; but measurement error actually impacts on its estimate with incidence proportional to the square of $\rho_{Y1,Y3}$. It is therefore possible to obtain an indicator of measurement reliability by separating the part that the model attributes to the actual variation of the underlying quantity.

For time-invariant variables, with $\rho_{X1,X2}=\rho_{X2,X3}=\rho_{X1,X3}=1$, (19) yields $\lambda_Y = \sqrt{\rho_{Yt,Yt+1}}$.

As noted, the index is based on the hypothesis that two independent first-order autoregressive models are a good approximation of the data-generating process. If this assumption does not hold, i. e. if a direct effect of X_1 on X_3 exists, the specification remains as described in (15), (16) and (17), but we have $\beta_{31}=\beta_{21}\beta_{32}+\beta^*_{31}$, where β^*_{31} is the regression coefficient relating X_1 and X_3 in the model including X_2; (20) becomes

$$\frac{\rho_{Xt-1,Xt}\,\rho_{Xt,Xt+1}}{\rho_{Xt-1,Xt+1}}+\frac{\rho^*_{Xt-1,Xt+1}\sqrt{(1-\rho^2_{Xt-1,Xt})(1-\rho^2_{Xt,Xt+1})}}{\rho_{Xt-1,Xt+1}}=1 \qquad (23)$$

This equation allows us to draw some general conclusions on what happens when the assumption of independent AR(1) processes is violated.

If we denote by ξ the term $\frac{\rho^*_{Xt-1,Xt+1}\sqrt{(1-\rho^2_{Xt-1,Xt})(1-\rho^2_{Xt,Xt+1})}}{\rho_{Xt-1,Xt+1}}$, we can write

$$\lambda^2_{Y\,AR(1)}=\frac{\rho_{Yt-1,Yt}\,\rho_{Yt,Yt+1}}{\rho_{Yt-1,Yt+1}}=(1-\xi)\lambda^2_{Y\,AR(2)} \qquad (24)$$

The Heise index measured under the AR(1) hypothesis is a distorted estimate of the reliability value that we would have if we took AR(2) processes into account. Since the partial and the simple correlation coefficients are usually positive or null if dealing with strongly persistent phenomena such as income and wealth, we can say that usually $0<(1-\xi)<1$; applying the base Heise method if the underlying data structure is AR(2) yields reliability estimates that are biased downwards[14].

4. The evaluation of data reliability: some evidence

4.1 Reliability of time-invariant quantities

Socio-demographic features that are either time-invariant (such as sex or year of birth) or subject to small changes only (such as educational qualification) are repeatedly measured on panel units. The study of discrepancies in reported values can shed some light on measurement error in the survey.

A number of inconsistencies emerge, even for the simplest questions. For example, 1.79% of respondents declared a different gender in 1989 and 1991; this percentage is stable if we compare the 1991 and 1993 waves, and it decreases in subsequent years, down to approximately 0.3% in recent times. The analysis of individual cases shows that 3 out of 4 times the error concerns young children, of whom no features other than the basic

[14] It is not easy to derive an unbiased estimator for the AR(2) case; it is not possible to obtain (22) by substitution, since the observed $\rho^*_{t+1,\,t-1}$ includes the very measurement error that we want to isolate. One possible solution is the correction of the Heise index by estimating the ξ component with instrumental variables.

demographics are recorded in the survey. The tendency for the misclassification rate to diminish with time is explained by the fact that from 1993 on a greater effort was made to avoid discrepancies; the introduction of CAPI in 1998 fortified the attempt with automatic consistency controls. Birth dates of respondents also show a small number of misalignments, again decreasing with time. The province of birth varies in 2% of cases; a slight increase in misclassifications in recent years is probably due to the introduction of new provinces (Table 4).

Table 4

INCONSISTENCIES IN CONTIGUOUS WAVES, 1989-2000

(percentages)

Waves	Sex	Year of birth	Place of birth
1989-1991	1.8	4.5	1.6
1991-1993	1.9	3.1	1.6
1993-1995	0.3	1.2	1.8
1995-1998	0.3	1.2	2.5
1998-2000	0.1	1.7	2.7

Some error-related problems emerge for educational qualifications, too. Since the level of education may actually change, of course – especially for the younger cohorts – it is necessary to study different age groups separately (Table 5). For the whole panel sample, 80% to 90% of the respondents declare the same qualification in two subsequent surveys. About 10% has qualifications that increase by one level in two years, a plausible event which is, reasonably, more frequent in the comparison between the 1995 and 1998 waves (the interval between them is larger than normal). An average 3-4% of respondents go from a higher qualification to a lower one, which implies that at least one of the two answers is wrong. Finally, very few declare that their educational attainment has jumped ahead by more than one level in two years, which is sometimes possible, sometimes not, depending on the situation. Focusing on the respondents aged 40 and older at the time of the interview, who have normally completed their education, the incidence of proven errors (diminishing qualifications) is more or less constant.

Table 5

EDUCATIONAL QUALIFICATIONS IN CONTIGUOUS WAVES, 1989-2000
(percentages)

Waves	All household members				Household members over 40			
	Same qualification	Qualification increasing by one level	Qualification decreasing	Qualification increasing by more than one level	Same qualification	Qualification increasing by one level	Qualification decreasing	Qualification increasing by more than one level
1989-1991...	87.2	9.3	3.1	0.5	92.8	2.8	4.0	0.4
1991-1993...	83.9	9.9	5.8	0.4	86.5	4.9	8.5	0.2
1993-1995...	86.8	11.3	1.7	0.2	94.0	3.2	2.5	0.3
1995-1998...	81.3	15.0	3.2	0.5	88.7	6.0	5.1	0.2
1998-2000...	88.4	9.4	1.9	0.2	93.4	3.9	2.6	0.1

It seems that this variable is more exposed to incorrect reporting than the socio-demographic traits assessed above. This could be an effect of response by proxy, if we assume that household members do not have precise information on one another's schooling. Moreover, the question could be perceived as sensitive, if people think that their qualifications are somehow too low.

Another feature that can be analyzed in order to gain insight on the reason why discrepancies arise is the type of high school diploma that respondents hold. Even if it only concerns a part of the panel sample (1,969 high school graduates), it is time-invariant: any reported difference can be safely labeled as an error. If we compare the 1998 and 2000 waves, we find that about 25% of the respondents report two different high school degrees. The transition matrix (Table 6) shows that almost 40% of inconsistencies arises between different types of trade schools, professional and technical.

Bachelor's degrees only change with time in 14.3% of cases; this might be caused by the proposed taxonomy, which includes a residual catch-all option (Table 7). A respondent may overlook the fact that her major is included in one of the first nine categories during one wave and then locate it correctly in the next wave, thus switching the answer from "other" to the right class. In addition to that, some persons may actually hold more than one university degree.

Table 6

TYPE OF HIGH SCHOOL DEGREE, 1998-2000

(percentages)

1998 \ 2000	A	B	C	D	E	F	Total
A. School for professional studies	3.3	4.7	0.9	0.1	0.4	0.3	9.8
B. Technical school	5.3	41.0	1.5	0.3	1.2	0.7	50.0
C. High schools specialized in classical, scientific or language studies	0.4	1.9	16.1	0.2	0.7	0.1	19.5
D. Art schools and institutes	0.2	0.3	0.6	2.0	0.1	0.0	3.2
E. Teacher training school.......................	0.6	1.1	1.0	0.0	11.6	0.0	14.3
F. Other	0.8	1.3	0.3	0.1	0.6	0.2	3.2
Total	10.5	50.4	20.4	2.7	14.6	1.4	100.0

Consistent answers (sum of diagonal elements): 74.3%.

Table 7

TYPE OF UNIVERSITY DEGREE, 1998-2000

(percentages)

1998 \ 2000	A	B	C	D	E	F	G	H	I	J	Total
A. Mathematics, physics, chemistry, biology, sciences, pharmacy...........	8.5	0.6	0.1	0.1	0.0	0.6	0.1	0.0	0.4	1.3	11.8
B. Agricultural or veterinary sciences .	0.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.3
C. Medicine and dentistry....................	0.3	0.3	5.8	0.4	0.0	0.0	0.0	0.1	0.3	0.1	7.3
D. Engineering....................................	0.0	0.0	0.0	9.1	0.0	0.1	0.1	0.1	0.0	0.0	9.5
E. Architecture or city-planning...........	0.0	0.0	0.0	0.0	4.0	0.1	0.0	0.0	0.1	0.0	4.3
F. Economics or statistics	0.1	0.0	0.0	0.0	0.0	8.9	0.0	0.3	0.0	0.6	9.9
G. Political science, sociology.............	0.1	0.0	0.0	0.0	0.0	0.0	2.2	0.0	0.0	0.3	2.6
H. Law...	0.0	0.0	0.0	0.0	0.0	0.3	0.0	10.8	0.1	0.4	11.7
I. Arts, philosophy, languages	0.7	0.0	0.3	0.0	0.1	0.4	0.1	0.4	28.4	0.6	31.1
J. Other	0.9	0.3	0.3	0.1	0.0	0.6	0.0	0.0	1.4	6.8	10.4
Total..	10.7	2.4	6.5	9.8	4.2	11.1	2.6	11.8	30.8	10.1	100.0

Consistent answers (sum of diagonal elements): 85.7%.

Response errors may be more frequent if the question itself is ambiguous or if the response options can be interpreted variously. For example, the answers given in 1998 and in 2000 on the location of the household's dwelling of residence (city center, suburbs, between the center and the suburbs etc.; Table 8) match only in 54.6% of the cases, probably because the classes are not precisely defined[15]. The same rules apply to the question concerning the quality of the neighborhood (upscale, rundown etc.), where the matching answers are 69.1% of the total (Table 9).

Table 8

LOCATION OF DWELLING OF RESIDENCE, 1998-2000

(percentages)

1998 \ 2000	Isolated area, countryside	Hamlet	Town outskirts	Between outskirts and town center	Town center	Other	Total
Isolated area, countryside	3.4	0.9	1.4	0.5	0.2	0.0	6.4
Hamlet...........................	0.8	3.0	2.0	0.5	0.1	0.0	6.3
Town outskirts	1.6	1.8	15.2	9.3	1.3	0.2	29.5
Between outskirts and town center....................	0.3	0.6	7.8	15.5	6.0	0.1	30.2
Town center....................	0.2	0.6	2.1	6.5	17.5	0.1	27.1
Other	0.1	0.0	0.2	0.2	0.1	0.0	0.6
Total	6.3	6.9	28.8	32.5	25.2	0.3	100.0

Consistent answers (sum of diagonal elements): 54.6%.

Table 9

QUALITY OF THE NEIGHBORHOOD OF RESIDENCE, 1998-2000

(percentages)

1998 \ 2000	Upscale	Neither upscale nor rundown	Rundown	Total
Upscale	12,2	13,7	0,5	26,4
Neither upscale nor rundown	11,0	55,8	2,1	68,9
Rundown	0,4	3,2	1,2	4,8
Total	23,6	72,7	3,7	100,0

Consistent answers (sum of diagonal elements): 69.1%.

In the case of continuous quantitative variables, the inconsistencies can be represented through a frequency distribution of the distance between answers given in two different waves. For example, the floor area of the dwelling of residence differs by less than 5 square meters in 44.5% of cases; the frequency decreases as the absolute value of the difference between two recorded values goes up (Figure 3; note that the data only concern respondents who didn't move or incur extraordinary renovation expenses between the two survey waves).

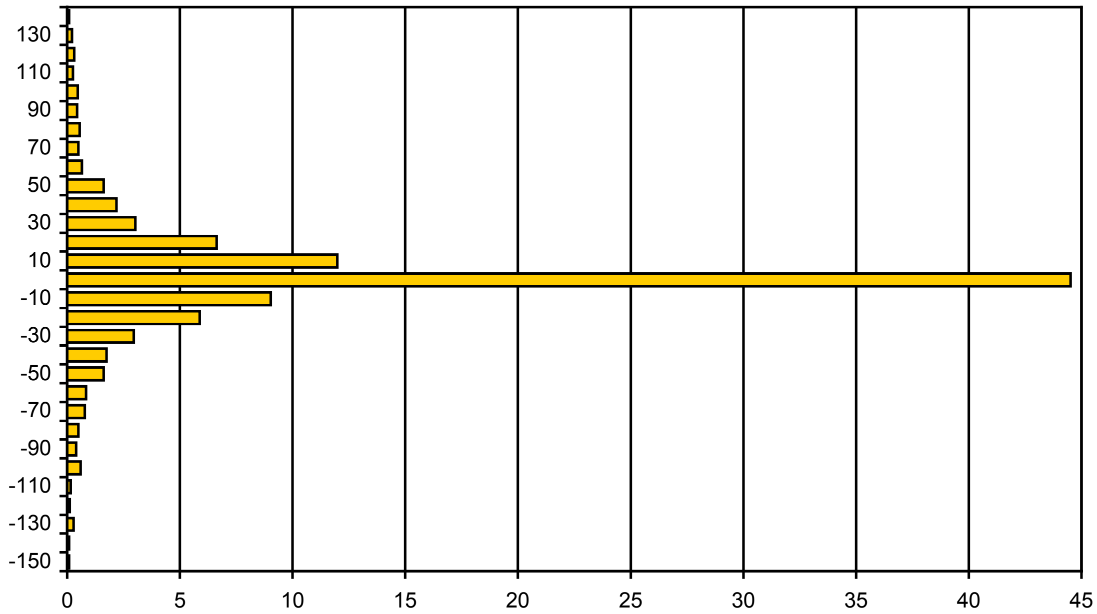
[15] The comparison has been carried out only on households that didn't move between 1998 and 2000. The "true" class of a dwelling is not necessarily time-invariant; changes, if any, should nevertheless only affect a small minority.

The linear correlation coefficient can be used as a reliability measure for time-invariant continuous variables; it corresponds to the square of the index λ^2. If no measurement error existed, it should equal 1; the less precise the data collection process, the greater the distance between ρ and 1 . For the surface area of the primary dwelling, it equals 0.65.

The correlation coefficient is lower for the construction year of the dwelling (ρ=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[16].

Figure 3

DISTRIBUTION OF DIFFERENCES IN THE REPORTED FLOOR AREA OF DWELLING OF RESIDENCE, 1998-2000
(percentages)



Another information that is affected by inconsistencies is the starting year of the respondents' working life. The usual recall problems are aggravated by a degree of ambiguity in the question: it is not clear whether occasional jobs or training periods should be included or not. Out of 5,117 individuals who answered the question both in 1998 and 2000, 46.5% gave answers that don't match; linear correlation is 0.64.

---

[16] The existence of renovations is, unfortunately, documented for the year 2000 only; a correct comparison would require data for 1999 too.

4.2 Reliability of time-varying quantities

Table 10 presents the Heise reliability index for the main variables collected in the 1995, 1998 and 2000 waves of the SHIW (Table 10)[17 18 19]. On a macro-aggregate level, net income and net wealth (Heise index: 0.82) seem to be more reliable than consumption (0.69)[20].

The income components that show most reliability are pensions and wages; both Heise indexes are around 0.95. Fringe benefits such as the right to drive a company car, on the contrary, are not recorded as precisely (0.41): probably it is not easy to express their monetary value. Data on self-employment and capital income are collected with less precision (the Heise indexes are, respectively, 0.74 and 0.72). Serious problems arise with information on depreciation (0.48) and distributed profits (0.35). Expenditure on food seems to show greater reliability (0.80) than consumption as a whole.

The Heise indexes for wealth items are quite heterogeneous. While real estate is surveyed quite well (0.80, with 0.96 for primary housing), valuables do not perform as satisfactorily (0.47); it might be hard to state the value of objects that are not currently on the market, especially when the price of acquisition is also unknown because they were inherited or received as gifts.

The index for financial assets as a macro-aggregate is 0.68. Government securities appear to be measured better than deposits and other securities (respectively 0.74 against

---

[17] The results presented in this section were obtained from the micro data of the historical archive, which includes imputed values. This implies that the reliability measure is referred to both collection and preliminary processing of information.

[18] Reliability indexes calculated for different sets of three waves (1989-1991-1993; 1991-1993-1995; 1993-1995-1998) are close to the ones presented in Table 10. Similar results are obtained by estimating Heise indexes on the basis of Spearman rank correlation coefficients, which are not as strongly affected by the presence of outliers as are the Pearson coefficients. Reasonably, the indexes for relatively unreliable quantities are also less stable: the only exception is consumer durables, always showing very low indexes. Probably the AR(1) model is not a satisfactory formalization of the data generating process for durables, since they are bought or renovated irregularly.

[19] The ranking of Heise indexes does not change even if we use, where necessary, the IV correction proposed for AR(2) processes. The direct application of IV methods for univariate regressions, when reasonably applicable (for example, when regressing consumption on income), also yields results that are aligned with Table 10.

[20] In order to identify the variables for which the assumptions are more likely to be violated, Heise suggests the comparison of $\rho_{41}\,\rho_{32}$ and $\rho_{31}\,\rho_{42}$, which can be calculated if we have four waves; if the AR(1) models are a good approximation of reality, the two quantities should be very close. In the SHIW, they very often are; significant differences exist for valuables and, to a lesser extent, family-owned businesses. Where income components are concerned, the largest discrepancy emerges for distributed profits.

0.38 and 0.64)[21]. Government bonds are perceived as not exposed to market fluctuations, since most holders do not sell them before their maturity date; in contrast to shares and mutual funds, respondents normally declare the face value of the bond, which is easy to remember. Deposits are measured with lower precision because their high degree of liquidity may induce memory problems.

Table 3

HEISE RELIABILITY INDEX FOR THE MAIN SURVEY AGGREGATES, 1995-1998-2000

Aggregate	Heise index
Income	
Net disposable income	0.82
Payroll income	0.94
Net wages and salaries	0.95
Fringe benefits	0.41
Pensions and net transfers	0.94
Pensions and arrears	0.95
Other transfers	0.76
Net income from self-employment	0.74
income from self-employment	0.79
depreciation	0.48
distributed profits	0.35
Net income from capital	0.72
income from buildings	0.67
income from financial assets	0.72
Wealth	
Net wealth	0.82
Real wealth	0.79
real estate	0.86
dwelling of residence	0.90
family-owned businesses	0.56
valuables	0.47
Financial wealth	0.68
deposits	0.38
government securities	0.74
other securities	0.64
Debts	0.54

Aggregate	Heise index
Consumption and savings	
Consumption	0.69
non-durables	0.69
expenditure on food	0.80
durables	0.27
Savings	0.61
Other aggregates	
Stock of durables	0.43
means of transport	0.89
furniture	0.23
Cash	0.57
Dwelling of residence	
Owners	
surface area	0.84
value	0.84
construction year	0.78
year of acquisition	0.83
imputed rent	0.74
Non-owners	
surface area	0.73
value	0.82
construction year	0.83
years of residence	0.96
rental rate	0.96

[21] A high value of the reliability index does not exclude problems such as the bias deriving from under-reporting; the Heise coefficient does not change if households systematically withdraw information on a part of their assets.

The measurement of debts appears to be quite unreliable (0.54). This applies to consumer durables as well (0.43), probably because the category encompasses many different types of goods, each of which induces different recall difficulties. Means of transport are an exception to this tendency (0.89), since information on the market value of used cars is widely available and known.

Finally, the value of primary housing is more reliable for the households that own than for those that rent; vice versa, actual rental rates are measured with greater precision than imputed ones.

5. Explanatory models for measurement error

The following paragraphs present models that aim to explain errors and inconsistencies in the survey data on the basis of fieldwork, interviewer and respondent features. Now, rather than seeking to quantify the incidence of measurement error, we want to find the reasons for it[22].

Paragraph 5.1 illustrates a model for the analysis of coding mistakes (e. g. wrong order of magnitude of a quantity); studying this problem is especially useful for an assessment of the interviewer's role in the determination of data quality, since the interviewer alone is responsible for erroneous coding.

During the preliminary editing phase, which precedes the production of statistics, the data undergo a number of quality controls. In many occasions, these controls lead to verification procedures which involve the examination of paper questionnaires (if available) or discussions with the interviewers. It would be theoretically possible to contact respondents again in order to remedy inconsistencies; but, as this is costly and time-consuming for the company in charge of the survey and the respondents alike, the actual strategy used is often different. If the discrepancies can be solved beyond reasonable doubt by looking at other sections of the questionnaire, the data is modified accordingly. This is

---

[22] A large part of the literature (e. g. Fabbris, 1989) claims that each interviewer induces an idiosyncratic distortion in answers, but the average bias is assumed to be null. If interviews were assigned casually, it would be possible to estimate the loss in precision caused by interviewers or by specific fieldwork features. We cannot assume casual allocation of assignments for the SHIW, because there is a strong correlation between the area in which an interviewer operates and respondent features. Moreover, this approach does not shed light on which features of data collectors actually affect the response variance; this is the reason why we prefer to study this problem with regression models.

typically the case, for instance, with real estate values expressed in millions of lire instead of thousands of lire; by comparing the declared worth with surface area and asset type it is easy to make the necessary correction. On the other hand, when the editing required is not so clear and the (presumed) inconsistency appears serious, households are re-contacted; if this is not possible, the answers are left as they are.

This approach reflects an interest in caution; anomalies are edited out of the data base only if they can be safely assumed to be errors, and inconsistencies are rectified only when the values are certainly wrong, and they can be univocally replaced with correct ones. Such caution avoids forced "normalization" of micro data, i. e. replacement of information describing uncommon but true situations with numbers that portray standard occurrences. Researchers using survey information are left with the responsibility of deciding how to treat anomalies, based on the specific features of their analysis.

It seems evident from what has been said so far that a study of the preliminary editing process can shed light on measurement error issues, although there are known limits to the insight that can be obtained from such an exercise. The frequency of editing actions remains an imprecise indicator of the incidence of measurement error on the survey. As stated above, these actions are carried out only when an item can be safely considered wrong; some problems are therefore left undocumented. Since interviewer mistakes are easier to catch than mistakes by respondents, a study of the preliminary editing phase is more helpful in relating interviewer features to errors than in explaining why households represent their economic situation incorrectly.

Section 5.2 presents inconsistencies in panel data for some socio-demographic variables and for income, discussing the features that often accompany them. Differently from the analysis of the editing process, to do this we need two or more waves. The 1998 and 2000 surveys have been selected, and discrepancies have been related to socio-demographic coordinates of respondents, interviewer features and fieldwork details[23].

---

[23] We could not find significant effects of interviewer and fieldwork features on average income. It is therefore possible to explain answer variability on the basis of such features as they were in each wave, without having to use their variations as additional controls.

It must be noted that information on interviewers is available for 2000 only. Since inconsistencies are generated by errors in either of the two waves, this specific lack in the data needs to be briefly discussed.

In a regression model, the omission of significant variables – such as interviewer features in 1998 – introduces a bias in the estimated coefficients; the extent of the bias depends on the correlation between the variables omitted and those included. In this specific case, since the two waves were carried out by different companies, it is reasonable to assume that these correlations are equal to zero, controlling for localization[24]. The marginal effects of each variable included in the model are therefore estimated without bias, even though they do not attain minimum variance as they would if the model were specified exactly.

5.1 Explanatory models: role of the interviewer

Keeping in mind the limits set out above, we now present some results concerning the preliminary editing process for the 2000 wave.

Two types of error are assessed. Firstly, we look at mistakes in the units of measurement; in some cases, answers are quite obviously given in millions of lire despite being requested in thousands of lire. Secondly, time-span errors are studied; in some cases, monthly incomes are declared when annual ones are requested and viceversa.

Issues related to units of measurement are mainly observed in the value of housing capital; time-span errors emerge on the income of employees and pensioners. These quantities are especially suitable for a preliminary study; the question on the value of the dwelling of residence is asked of every household in the sample, and a good share of the respondents are employees and pensioners, who are also not as likely to under-report as the self-employed, because they normally receive their income already net of tax. The focus on these two types of income decreases the representativeness of our analysis, but it also allows us to identify the errors in the data with ease and to explain them with simple models.

---

[24] As illustrated above, the distribution of interviewer features is conditioned by localization. This induces positive correlation between features in different periods, but it is possible to eliminate its effects by including a geographical dummy in the regressions.

Table 11 gives some descriptive statistics on the editing actions taken on the three variables cited; a first striking fact is that error concentration is high. If the interviewers are ordered by number of errors found in their work, the last quartile appears responsible for a share of editing actions that ranges between 78.6 and 88.6 per cent. The Gini coefficient for the number of errors ranges between 0.34 and 0.48.

The correlation coefficient between the percentage of errors and the number of interviews is always negative; the company in charge of the survey probably operates some form of control, giving more assignments to the better interviewers[25].

Table 4

ERRORS ON SOME SURVEY VARIABLES, 2000

(units, percentages)

	Value of dwelling of residence	Income (employees)	Income (pensioners)
Errors	238	152	534
Records	8,001	6,553	6,175
Error rate	3.0	2.3	8.6
Interviewers with no errors	59.8	69.8	63.1
Interviewers with 1 error	22.0	15.3	14.0
Interviewers with more than 1 error	18.2	14.9	22.9
Error concentration	0.34	0.37	0.48
Share of errors accruing to interviewers with the worst performances (25 per cent with highest number of errors)	78.6	88.7	86.8
Correlation between error rate and number of interviews	–0.12	-0.13	-0.003

This evidence also suggests that the skill of each interviewer is indeed observable; there is no *a priori* knowledge of how many corrections will be made on each individual interview, but the distribution of assignments seems to be consistent with *ex post* measures of precision.[26]

[25] The interpretation of this information is not straightforward. While 59.8% of interviewers commit no mistakes, just 50.6% of the questionnaires require no editing actions. This indicates that the best interviewers typically carried out a number of assignments that is below the general average. But the number of questionnaires with a single error is 25.5% of the total (22.0% of the interviewers); 10.4% of interviews have to be edited twice (9.0% of interviewers).

[26] Some reflections on efficient methods of interviewer selection and supervision can be found in Fowler (1991) and, with specific reference to new interviewing technologies, in Nicholls, Baker and Martin (1997).

Table 12 presents the results of a logistic regression run on edited and unedited records concerning the value of primary housing, the income of employees, and the income of pensioners. Only the first 60 interviews carried out by each interviewer are considered (interval of one standard error around the mean number of assignments), in order to limit the effect that last-minute interviews – typically assigned to top interviewers, hence not really representative – have on our estimates. Therefore, 6,467 questionnaires are considered (80.8 per cent of the sample), corresponding to 722 edited records out of 924 (78.2 per cent of the total).

The dependent variable is a dummy, set at 1 if the record has been corrected, 0 otherwise; the vector of independent variables encompasses the main features of interviewers and some fieldwork details, such as the use of CAPI and the length of the interview.

The respondent might have a role in determining mistakes of the type we are now studying, but this role appears to be junior high: mismatch in the measurement units, failure to clarify the time horizon relative to each question, lack of consistency checks are signals of incorrect interviewer behavior.

The results of our logistic regression are sufficiently stable. The probability of recording data that will subsequently require correction seems to be influenced by both interviewer and fieldwork features.

Predictably, professional interviewers are less inclined to go wrong; also, as the fieldwork progresses experience is gained and the frequency of errors diminishes. The non-professional interviewers who manage to obtain high response rates are also more precise.

Figure 4 describes the impact of interview length on accuracy. If too little or too much time is taken, data quality suffers. Normally, about one hour is needed to collect information. If the interviewers are excessively fast, they are probably not paying enough attention. Remarkable slowness might be a signal of interaction problems; moreover, fatigue can affect both interviewers and respondents if the interview exceeds average length.

Table 5

PROBABILITY OF WRONGLY RECORDING AT LEAST ONE ANSWER, 2000

(logit estimate)

	Coefficient°
Intercept	36,3734**
North	-0.0449
Center	-0.1954
Municipality: up to 20,000 residents	0.3212**
Municipality: between 20,000 and 40,000 residents	0.1976
Municipality: between 40,000 and 500,000 residents	0.2900
Paper questionnaire	-1.2401***
Interview length	-0.0403***
Interview length squared	0.0003***
Interviewer assessment of the general psychological climate during the interview °°	-0.1159***
Progressive number of the interview in the interviewer's portfolio	-0.0195***
Interviewer: previous SHIW waves	-0.0454
Interviewer: birth year	-0.0182**
Interviewer: male	-0.2056
Interviewer: junior high school degree	-0.4819
Interviewer: high school degree	-0.4948**
Interviewer: resident in a province different from the respondent	0.4083**
Interviewer: response rate	0.5137
Interviewer: non-professional	0.8381**
Non-professional interviewer: response rate	-1.2495**

*** Significant at the 1% level **Significant at the 5% level *Significant at the 10% level.
°Significance levels take into account intraclass correlation coefficients for each interviewer
°° Score expressed out of ten.

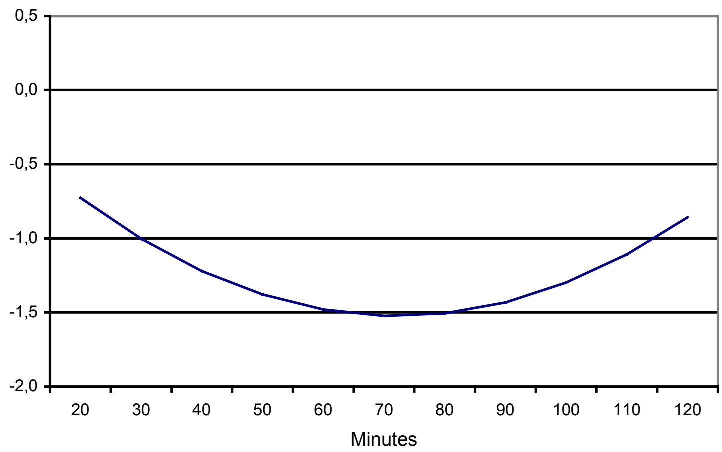
Complementary results are obtained when studying the impact of the psychological climate in which the interview is carried out, as assessed by the interviewer; if tension arises, e. g. in the cases where some effort is needed to overcome reticence, more errors appear.

Where demographic traits of the interviewers are concerned, it seems that young males with high school degrees are less likely to be mistaken.

The risk of errors increases if the data collector is operating in a province different from his own. This may depend on a number of factors; first of all, "away" interviews tend to be shorter; interviewers are in a hurry and concentrate less. The psychological climate also tends to worsen.

Figure 4

EFFECT OF INTERVIEW LENGTH ON THE PROBABILITY OF WRONGY RECORDING AT LEAST ONE ANSWER
(logit estimate)



On average, errors are less frequent when a paper questionnaire is used. It must be pointed out that the CAPI program is very efficient in monitoring the interview flow; consistency controls, on the other hand, are limited, because it is not possible to exclude most unusual answers *a priori*. The software therefore only asks the user to confirm recorded anomalies, and it is only as accurate as the interviewers. Moreover, the CAPI interface proposes a sequence of screens and it is not convenient to go back to the previous ones to verify internal consistency; a paper questionnaire allows simultaneous vision of all answers, which is an advantage for quality control. The problem worsens if interviewers are not adequately trained using CAPI, as noted by Couper, Hansen and Sadoski (1997). This effect seems to offset the positive impact of automated verification mechanisms.

The superior quality of data collected via paper questionnaire does not hold for all variables; in particular, as we will see shortly, the CAPI method performs better on the items for which specific controls are implemented.

5.2 Explanatory models: role of the interviewer and role of the respondent

5.2.1 Educational qualifications

In this section we present the results of a logistic regression on the 9,473 individuals who declared their educational qualifications both in 1998 and 2000 (Table 13).

The dependent variable is a dummy that has value 1 if an inconsistency is reported[27]: a qualification that is higher in 1998 than in 2000, and a qualification that is higher by more than one level in 2000 than in 1998. The cases of one-level rise (e. g. primary/junior high school) were not labeled as erroneous.

Table 6

PROBABILITY OF FINDING INCONSISTENT ANSWERS ON EDUCATIONAL QUALIFICATIONS, 2000

(logit estimate)

	Coefficient°
Intercept	47.2256**
Respondent: male	-0.2512*
Respondent: no formal education	-0.5377*
Respondent: primary school degree	-0.5637***
Respondent: junior high school degree	0.1367
Respondent: high school degree	0.6588***
Respondent: birth year	-0.0182***
North	-0.1276
Center	-0.4366
Municipality: up to 20,000 residents	-0.3626
Municipality: between 20,000 and 40,000 residents	-0.3188
Municipality: between 40,000 and 500,000 residents	-0.2916
Paper questionnaire	0.3465*
Interview length	0.0005
Personal interview +	-0.2473**
Interview by proxy+	0.2694*
Interviewer assessment of the general psychological climate during the interview °°	-0.1755*
Progressive number of the interview in the interviewer's portfolio	-0.0070
Interviewer: previous SHIW waves	-0.0160
Interviewer: birth year	-0.0671
Interviewer: male	-0.0070
Interviewer: junior high school degree	0.4248
Interviewer: high school degree	-0.0604
Interviewer: resident in a province different from the respondent	0.2183
Interviewer: response rate	-0.0359
Interviewer: non-professional	-0.0319

*** Significant at the 1% level **Significant at the 5% level *Significant at the 10% level.
+ The base class is "unknown".
°Significance levels take into account intraclass correlation coefficients for each interviewer
°° Score expressed out of ten.

[27] It is useful to point out that the estimates discussed here do not refer (as the previous ones did) to the probability of wrongly recording an answer; they concern the probability of finding inconsistent answers. The controls refer to what was stated by respondents in 1998, even if there is no reason to believe that in the presence of a discrepancy the true answer is the one given in any specific year. The replacement of 1998 values with 2000 values does not produce significant differences in the results. This implies a substantial symmetry in the probability of finding inconsistencies with respect to the wave chosen as portraying the "correct" educational qualification. In other words, the probability of observing discrepancies for a given class is approximately the same if the class is studied in 1998 or 2000. This notwithstanding, the presence of measurement error in the independent variables should introduce a distortion on coefficients known as *attenuation bias*.

In this case, like others that will be discussed below, respondent and fieldwork features are likely to be useful in explaining the presence of discrepancies; on the contrary, interviewers do not seem to play a crucial role.

Male respondents appear to be more consistent in their answers; elderly people seem to be more inclined to report two different qualifications. This can be explained with recall problems (Pearson, Ross and Dawes, 1992). Those who were born before 1955 also experienced an educational system different from the current one: they could choose between junior high school and apprenticeship, and possibly they have trouble reconciling their experience with one of the response options, which refer to the present organization of schools.

This particular circumstance helps explain the concentration of inconsistencies on intermediate qualifications; those who had no formal education or completed primary school only tend to confirm in 2000 what they stated in 1998, while those who chose “junior high school” or “high school” are more exposed to confusion (the effect is statistically significant for high school only). Finally, even if the question clearly refers to the highest *attained* qualification, drop-outs at various levels may be uncertain in describing their situation.

If we look at fieldwork features and interviewer-respondent interaction, personal interviews are less exposed to discrepancies than the ones conducted by proxy. A good psychological climate reduces the chances of error; paper questionnaires are worse than CAPI, since the software actually points out inconsistencies for this particular question.

Previously outlined quality-improving factors, such as the experience gained in the course of a particular wave, as measured by the number of interviews already carried out, or the residence of the interviewer and the respondent in the same province, do not seem to be relevant in this case. The same goes for interviewer features, such as previous involvement in the SHIW: most coefficients have the expected sign, but they are not statistically significant.

5.2.2 Type of high school degree

The logit regression ran on the basis of answers provided by high school graduates interviewed in both 1998 and 2000 on their type of degree confirms that the discrepancies depend heavily on the degree itself (Table 14)[28]; those who reported graduation from a trade school in 1998 were more likely than the rest to change their answer in 2000. Degrees with a higher level of specificity seem to induce less confusion. Males turn out to be more consistent again; the remaining features do not seem to be significant. The general psychological climate is, again, correlated with a smaller probability of error. The interviewers who have a long record of SHIW waves, reside in the same province as the households surveyed, and hold a junior high school degree are less inclined to record inconsistencies.

Table 14

PROBABILITY OF FINDING INCONSISTENT ANSWERS ON THE TYPE OF HIGH SCHOOL DEGREE, 2000

(logit estimate)

	Coefficient°
Intercept	16.3527
Respondent: male	-0.2450*
Respondent: school for professional studies	1.1456***
Respondent: technical school	-1.0692***
Respondent: high school specialized in classical, scientific or language studies	-1.5465***
Respondent: art schools and institutes	-0.3923
Respondent: teacher training school	-1.2405***
Respondent: birth year	-0.0001
North	0.2935
Center	0.4309**
Municipality: up to 20,000 residents	0.0086
Municipality: between 20,000 and 40,000 residents	-0.0823
Municipality: between 40,000 and 500,000 residents	0.1400
Paper questionnaire	0.0953
Interview length	0.0026
Personal interview +	-0.2370***
Interview by proxy+	-0.2253
Interviewer assessment of the general psychological climate during the interview °°	-0.1644
Progressive number of the interview in the interviewer's portfolio	-0.0055
Interviewer: previous SHIW waves	-0.0693**
Interviewer: birth year	-0.0074
Interviewer: male	-0.0138
Interviewer: junior high school degree	-0.6487**
Interviewer: high school degree	-0.1859
Interviewer: resident in a province different from the respondent	0.2922*
Interviewer: response rate	0.1753
Interviewer: non-professional	0.0556

*** Significant at the 1% level **Significant at the 5% level *Significant at the 10% level.
+ The base class is "unknown".
°Significance levels take into account intraclass correlation coefficients for each interviewer
°° Score expressed out of ten.

[28] A substantial symmetry in the distribution of inconsistencies exists, as stated for educational qualifications.

5.2.3 First year of working life

A logit regression run on inconsistencies in the reported year in which respondents started working further highlights the role of respondent features in determining the probability of discrepancies (Table 15).

Table 15

PROBABILITY OF FINDING INCONSISTENT ANSWERS ON THE FIRST YEAR OF WORKING LIFE, 2000

(logit estimate)

	Coefficient°
Intercept	24.3073***
Respondent: male	-0.1843***
Respondent: birth year	-0.0091***
Respondent: no formal education	0.3370***
Respondent: primary school degree	0.0559
Respondent: junior high school degree	-0.2839***
Respondent: high school degree	-0.0869
Respondent: employee	-0.1304**
Respondent: self-employed	0.0618
Respondent: pensioner, former employee	-0.0315
Respondent: number of jobs held	-0.0075
North	-0.6519***
Center	-0.3123***
Municipality: up to 20,000 residents	0.1777**
Municipality: between 20,000 and 40,000 residents	0.1969**
Municipality: between 40,000 and 500,000 residents	0.6758***
Paper questionnaire	0.3201***
Interview length	0.0014***
Personal interview +	-0.1126***
Interview by proxy+	-0.0308
Interviewer assessment of the general psychological climate during the interview °°	0.1192***
Progressive number of the interview in the interviewer's portfolio	0.0025
Interviewer: previous SHIW waves	-0.0080
Interviewer: birth year	-0.0029
Interviewer: male	-0.0270
Interviewer: junior high school degree	-0.3900***
Interviewer: high school degree	-0.2093***
Interviewer: resident in a province different from the respondent	0.0742
Interviewer: response rate	0.5041**
Interviewer: non-professional	0.7525**
Non-professional interviewer: response rate	-0.8970**

*** Significant at the 1% level **Significant at the 5% level *Significant at the 10% level.
+ The base class is "unknown".
°Significance levels take into account intraclass correlation coefficients for each interviewer
°° Score expressed out of ten.

Once again, answers provided by males are more stable. The elderly face the usual recall problems; employees are better than the self-employed at remembering when they started working, probably because the concept itself is more formalized for them, and the initial date is more frequently recalled for reasons connected to wages, promotions and pensions.

Educational qualifications have an interesting effect on inconsistencies: those who are at the bottom and at the top of the qualification ladder are more exposed to errors. Where the unschooled are concerned, this can be explained by the usual difficulties in understanding the questions and interacting with the interviewers. University graduates, on the other hand, might have worked part-time while students, and they might be undecided as to whether they should consider these (often occasional) jobs as part of their working life or not.

It is also worth noting that in this case the geographical covariates, normally not significant, reveal more inconsistencies in the South and in small towns, possibly because the informal economy is more important there.

The interviewer-respondent interaction is again significant; interviews by proxy are less precise than personal ones, professional interviewers and the non-professional ones who obtain higher response rates are less exposed to error. A good interpersonal relationship between the interviewer and the family, as signaled by the psychological climate in which the interview is carried out, shows positive effects. Excessive length of the assignment produces the adverse consequences.

The CAPI technique reduces the probability of inconsistencies. Even if there is no specific control for this question, the initial working age is cross-examined with other variables, such as the year of birth, the number of years in which the respondent has paid pension contributions, and the year of retirement.

The effect of interviewer experience with the SHIW and residence in the same province as the respondent have the expected sign, but they are not statistically significant. Socio-demographic features of the interviewer, on the other hand, do not seem to explain inconsistencies at all, with the exception of educational qualifications: interviewers with junior high school degrees seem to perform better.

5.2.4 Income

The most important cause of measurement error for income data is under-reporting, which is the voluntary underestimation of earnings. The fundamental traits of this behavior have been discussed in several studies[29] and are found, on a qualitative level, in the indications that interviewers give on the credibility of respondents when they evaluate the questionnaire. According to these indications, income and wealth are underestimated more severely in the South, and when the head of the household is self-employed, poorly educated, or elderly (Table 16).The extent of under-reporting could be assessed using *ad hoc* analyses alone; in the context of this study, however, it is interesting to understand if fieldwork organization or interviewer features influence the variability of recorded incomes.[30]

The investigation into the causes of discrepancies in incomes reported for 1998 and 2000 has been carried out with two different models[31]. The first relates, via a linear regression, the absolute value of differences between the two waves to a number of controls that should catch "true" variations, and to the usual interviewer and fieldwork feetures. The coefficients yield a measure of the impact of each observed factor on the observed variability. The second is based on a logistic regression that models the probability of observing discrepancies greater than a fixed limit. In particular, assuming that the mistakes made by interviewers mostly appear in the tails of the distribution of differences, we create an indicator variable signaling percentage variations of income below the 5th and above the 95th percentile.[32]

Given that the results are robust, we only present the logistic regression since its outcome is less exposed to the influence of outliers. Note that here, in contrast to our earlier procedure, income data is studied after the preliminary editing, and it is hence already devoid of blatant inconsistencies.

---

[29] See: Cannari and D'Alessio (1990); Cannari and D'Alessio (1993); Cannari, D'Alessio, Raimondi and Rinaldi (1990); Cannari and Violi (1995); Brandolini (1999).

[30] As we stated above, some experiments have been carried out to evaluate the impact of interviewer and fieldwork features on the average levels of recorded income; no significant effects have been found.

[31] Since no information on interviewers is available for the 1998 wave, results have to be interpreted on the assumption of independence between interviewer features in the two periods.

[32] In order to eliminate the effect of changes in household composition, only the households with the same roster have been studied.

Table 16

INTERVIEWER ASSESSMENT OF THE CREDIBILITY OF ANSWERS ON INCOME AND WEALTH

(scores out of ten)

Feature (*)	Credibility of answers
Sex	
Male	7.7
Female	7.8
Age class	
Up to 30	7.8
31 to 40	7.9
41 to 50	7.8
51 to 65	7.7
Over 65	7.5
Educational qualification	
No formal education	7.1
Primary school degree	7.3
Junior high school degree	7.8
High school degree	8.0
University degree	8.1
Sector of employment	
Agriculture	7.1
Manufacturing	7.7
Government	8.1
Other sectors	7.8
No sector	7.6
Occupational status	
Employee	
Blue-collar worker	7.7
Office worker	8.1
Manager	8.3
Total	7.9
Self-employed	
Professional. sole proprietor	7.7
Other self-employed	7.3
Total	7.5
Not employed	
Pensioner	7.6
Other not employed	7.0
Total	7.6

Feature (*)	Credibility of answers
Number of family members	
1 member	7.6
2 members	7.7
3 members	7.8
4 members	7.8
5 members and up	7.5
Number of earners	
1 earner	7.5
2 earners	7.9
3 earners	7.8
4 earners and up	7.7
Income class	
Up to 10 thousand euros	7.0
10 to 20 thousand euros	7.5
20 to 30 thousand euros	7.9
30 to 40 thousand euros	8.0
40 thousand euros and up	8.1
Municipality size	
Up to 20.000 residents	7.9
20.000 to 40.000 residents	7.6
40.000 to 500.000 residents	7.7
500.000 residents and up	7.7
Geographical area	
North	8.1
Center	7.6
South and Islands	7.3
Total	**7.7**

(*) Individual features are referred to the head of household (main earner).

Source: Bank of Italy, 2002.

The experiment analyzed the differences between incomes reported by employees and pensioners in 1998 and 2000, corresponding to 3,244 households. As expected, a large part of the variability is related to socio-demographic features such as changes in the number of earners, gender, type of occupation, educational qualification, and area of residence (Table 17). Operational conditions and interviewer features do not seem to impact significantly on the discrepancies, except for the interviewer's experience and the general psychological climate in which the interview is carried out, both of which show the expected sign.

Variability in reported incomes therefore seems to depend on causes external to the survey process itself; the very tendency towards under-reporting could be a further cause of additional variance, if the underestimation is not systematic.

Table 17

PROBABILITY OF FINDING INCONSISTENT ANSWERS ON THE STARTING YEAR OF WORKING LIFE, 2000

(logit estimate)

	Coefficient°
Intercept	-30.8911
Respondent: male	-0.3287
Respondent: birth year	0.0117
Respondent: number of earners in the household	0.1602
Respondent: new earners (employees) in the household	1.0382***
Respondent: new earners (self-employed) in the household	1.3093***
Respondent: new earners (transfers) in the household	0.8781***
Respondent: household wealth below general median	1.0952***
North	0.3960**
Center	-0.1302
Municipality: up to 20,000 residents	-0.3370
Municipality: between 20,000 and 40,000 residents	-0.3141
Municipality: between 40,000 and 500,000 residents	-0.0822
Respondent: no formal education	-0.9134
Respondent: primary school degree	-0.1442
Respondent: junior high school degree	-0.2048
Respondent: high school degree	0.2445
Respondent: employee	-0.6571***
Respondent: self-employed	0.9803***
Respondent: new head of household	0.3150
Paper questionnaire	0.0711
Interview length	0.0015
Interviewer assessment of the general psychological climate during the interview °°	-0.1230**
Progressive number of the interview in the interviewer's portfolio	-0.0032
Interviewer: previous SHIW waves	-0.1164***
Interviewer: birth year	0.0028
Interviewer: male	0.1594
Interviewer: junior high school degree	0.0395
Interviewer: high school degree	0.2877
Interviewer: resident in a province different from the respondent	0.1274
Interviewer: non-professional	0.2247
Interviewer: response rate	-0.1329

*** Significant at the 1% level **Significant at the 5% level *Significant at the 10% level.
°Significance levels take into account intraclass correlation coefficients for each interviewer
°° Score expressed out of ten.

6. Conclusions

We have analyzed measurement errors affecting the most important variables of the Bank of Italy’s Survey of Household Income and Wealth (SHIW). The results are relevant, first of all, to researchers who use the survey, since data quality has to be taken into account. Also, knowledge of problems with the data is essential to improving the survey itself.

For time-invariant quantities, we evaluated the consistency of the answers given by panel units in various waves; for time-variant ones, such as income or wealth, we used the Heise (1969) model, which under mild regularity conditions can separate the actual change in a variable from measurement error on the basis of three or more subsequent waves.

We also examined the role played by fieldwork, interviewer and respondent features. Along with idyosincratic elements referred to specific questions, there are a number of common explanatory factors for discrepancies.

The main results are as follows:

- Inconsistencies arise for all questions, even those that are neither ambiguous nor difficult to understand: in this case, discrepancies amount to 2 or 3 per cent of the total. The number of errors decreases with time as greater attention was paid to avoiding them. Three fourths of the inconsistencies concern young children, surveyed only with respect to basic demographics.

- The number of inconsistencies increases when the question concerns information that might not be available to all family members, or is perceived as sensitive by the respondent, such as the type of high school degree or the level of educational qualification.

- When a question involves memory (e. g. the age at which a respondent started working) or when it does not specify how to treat certain situations (e. g. apprenticeship or occasional jobs), inconsistencies are the result of objective difficulties in determining the correct answer.

- Errors are more frequent when the response options are not precise enough (e. g, “center” or “suburbs” in the question about the location of one’s primary residence).

- The Heise reliability index, which measures the level of precision of the data (but does not catch systematic under-reporting), is higher for income and wealth (0.82) than for consumption (0.69).

- As to income, the data for employees and pensioners are the most reliable (around 0.95). Fringe benefits, on the contrary, are quite problematic, showing a Heise index of 0.41. Income from self-employment and capital are collected less precisely (respectively, 0.74 and 0.72).
- The consumption component that performs best is food (0.80).
- The Heise index for real estate wealth is 0.86; primary residence performs even better (0.90). Other wealth components, such as valuables (0.47), are more exposed to error, since it is not easy to evaluate items that are not currently on the market.
- Personal interviews contain fewer discrepancies than those conducted by proxy.
- When the CAPI software includes specific consistency controls, this helps to avoid discrepancies; when such controls are not present, the paper questionnaire is more precise. This is probably because it allows simultaneous view of the answers, where the electronic interface requires switching back and forth between screens. The problem is worse when the interviewer is not adequately trained in the use of the program.
- Interview length has an impact on accuracy; if it is too short or too long, data quality worsens. About one hour appears to be needed to complete the questionnaire. If the interview is much shorter, probably the interviewer didn't pay enough attention; if it is much longer, fatigue may set in. Moreover, long interviews probably reflect a difficult interviewer-respondent interaction.
- Professional interviewers (and non-professionals who are better at obtaining high response rates) have better results in terms of data quality; previous experience with the SHIW has similar results.
- Experience gathered during a wave, as measured by the number of interviews already carried out, improves accuracy. The last assignments of each interviewer, in fact, are on average significantly better than the rest.
- The risk of errors increases when the interviewer works in a province other than that of residence. This can depend on several factors. Controlling for the number of family members and income earners, "away" interviews are shorter than "home" interviews; possibly, time constraints intervene.

References

Banca d'Italia (2002), I bilanci delle famiglie italiane nell'anno 2000, a cura di G. D'Alessio, I. Faiella, Supplementi al Bollettino Statistico (nuova serie), Banca d'Italia, Anno XII n. 6.

Biemer, P., D. Trewin (1997), A Review of Measurement Error Effects on the Analysis of Survey Data, L. Lyberg et al. eds., Survey Measurement and Process Quality, Wiley, pp. 603-633.

Brandolini, A. (1999), The Distribution of Personal Income in Post-War Italy: Source Description, Data Quality, and the Time Pattern of Income Inequality, Giornale degli Economisti e Annali di Economia, n. 2, Vol.58, pp. 183-239.

Cannari, L., G. D'Alessio (1990), Housing Assets in the Bank of Italy's Survey of Household Income and Wealth, Income and Wealth Distribution, Inequality and Poverty - Proceedings, Pavia, Dagum and Zenga eds., Springer Verlag, pp. 326-334.

Cannari, L., G. D'Alessio (1992), Mancate interviste e distorsione degli stimatori, Temi di discussione, Banca d'Italia, n. 172, Giugno.

Cannari, L., G. D'Alessio (1993), Non-Reporting and Under-Reporting Behavior in the Bank of Italy's Survey of Household Income and Wealth, Bulletin of the International Statistical Institute - Proceedings of the ISI 49th Session, tome LV, four books, Isi, n. 3, Agosto-Settembre, pp. 395-412.

Cannari, L., G. D'Alessio, G. Raimondi, A.I. Rinaldi (1990), Le attività finanziarie delle famiglie italiane, Temi di discussione, Banca d'Italia, n. 136, Luglio.

Cannari, L., R. Violi (1995), Reporting Behaviour in the Bank of Italy's Survey of Italian Household Income and Wealth, Research on Economic Inequality, Vol. VI, JAI Press Inc., pp. 117-130.

Couper, M.P., S.E. Hansen, S.A. Sadosky (1997), Evaluating Interviewer Performance in a CAPI Survey, L. Lyberg et al. eds., Survey Measurement and Process Quality, Wiley, pp. 267-285.

D'Alessio, G., I. Faiella (2002), Non-response Behaviour in the Bank of Italy's Survey of Household Income and Wealth, Tema di discussione n. 462, Banca d'Italia.

Fabbris, L. (1989), L’indagine campionaria, La Nuova Italia Scientifica.

Fowler, F.J. (1997), Reducing Interviewer-Related Error Through Interviewer Training, Supervision, and Other Means, P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, S. Sudman eds., Measurement Error in Surveys, Wiley, pp. 259-278.

Groves, R.M., M.P. Couper (1998), How Survey Design Features Affect Participation, Non-response in Household Interview Surveys, Wiley, pp. 269-293.

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

Heise, D. (1969), Separating Reliability and Stability in Test-Retest Correlation, American Sociological Review 34 (1):93-101.

Lord, F. M., M. R. Novick (1968), Statistical Theories of Mental Test Scores, Addison-Wesley.

Nicholls, W., R. Baker e J. Martin (1997), The Effect of New Data Collection Technologies on Survey Data, L. Lyberg et al. eds., Survey Measurement and Process Quality, Wiley, pp. 221-244.

Pearson, R.W., M. Ross e R.M. Dawes (1992), Personal Recall and the Limits of Retrospective Questions in Surveys, J. M. Tanur ed., Questions about Questions, Russell Sage, pp.65-94.

Silverman, B.W. (1986), Density Estimation, Chapman & Hall.