doi.org/10.26398/IJAS.0032-013

MEASUREMENT ERRORS IN SURVEY DATA AND THE ESTIMATION OF POVERTY AND INEQUALITY INDICES

Giovanni D'Alessio

Bank of Italy, Economic Research and International Relations, Rome, Italy
Via Nazionale 191 (Villa Huffer), 00184 ROMA

Abstract. *This paper firstly provides simple tools for evaluating the incidence of measurement error affecting the main variables collected in surveys on consumption. The assessment is carried out on two surveys that provide both diary and panel data. Diary data can be employed to obtain reliability coefficients for time-invariant variables. When variables vary over time, an estimation of the incidence of measurement error on the total variance can be obtained by applying models that allow the decomposition of observed variability into true dynamics and noise. Evaluations are also conducted on the basis of the internal consistency criterion. Finally, some methods for estimating the impacts of measurement errors on poverty and inequality analysis are also discussed.*

Keywords: *Inequality, poverty, survey data, measurement errors, reliability.*

1. INTRODUCTION[1]

In consumption surveys, the discrepancy between recorded and 'true' data, which originates from the response or from oversights in the processing phase before estimation, is assumed to be important.

The survey design in all its parts may have an impact on survey responses. Aggregated variables such as 'household consumption' are usually derived by summing up dozens of items, collected by asking many household members several questions. In some cases, questions can be asked ambiguously or face limitations due to the cognitive processes of the respondent: people may not actually know the exact answer to the questions they are asked, especially in cases where response by proxy is allowed, and answers on quantities tend to be rounded up or down. Moreover, retrospective questions mean recalling events

Email: giovanni.dalessio@bancaditalia.it

[1] I thank the two anonymous reviewers whose comments and suggestions helped improve and clarify this manuscript. The opinions expressed in this article are mine alone and do not necessarily represent the views of the Bank of Italy.

of the past, while hypothetical ones require some abstract reasoning that may generate uncertain answers. All the above aspects may interact among themselves and with other factors affecting the quality of survey data. Interviewer behaviour, for example, can be 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. More general aspects, such as the motivation of respondents and their willingness to give their time and effort to a survey, should also be assumed to influence data quality.

Assessing the origin and the amount of measurement errors in survey data is important, due to the impacts they may have on inequality and poverty estimates. According to classical hypotheses, errors add noise and tend to inflate the variance and the tails of the distribution, thus boosting inequality and poverty indices.

In this paper, we will first focus on the tools enabling the assessment of the magnitude of measurement errors in the main variables collected in consumption survey data. We will show how measures of reliability[2] can be estimated using survey data, and will discuss some typical drivers of measurement errors in consumption surveys. This approach does not require the availability of true data to make a comparison with survey data, although it confines the analysis to those cases where a hypothesis of classical measurement errors holds, at least approximately. In doing this, we also resort to statistical tools developed in the psychometric literature, customizing their application for the field of consumption surveys.

Information on the share of measurement error contained in survey variables is useful for data producers, who may find a tool for discussing and improving the data collection process. On the other hand, it is equally important that users are aware of the amount and the expected effects on estimates of measurement errors affecting data.

The methods for quantifying measurement errors have already been described in the literature, although quite rarely in the field of consumption surveys. The techniques employed for obtaining adjusted estimates are also

[2] The reliability of a measure denotes the variability of the estimates over repeated trials and in the same approximate conditions. It is different from the accuracy of a measure, which implies both a small variability of estimates and a closeness to the true value (Hand et al., 2001).

common in regression analysis[3] but quite rare in poverty and inequality analysis.

This paper is characterized by a unified presentation of these methods and by their application to consumption surveys, which often use diary data, adopt panel sampling schemes and collect information on many correlated variables, thus offering specific opportunities for the analysis of measurement error.

The study was conducted using data from two surveys conducted in Tanzania. The tools proposed are of course general and could easily be extended to similar surveys.

The paper is structured as follows. Section 2 presents a short review of the existing literature on measurement errors in surveys on consumption, income and wealth. Section 3 shows the statistical tools that can be used for evaluating the degree of reliability of collected survey data. Section 4 describes some examples of how the tools can be employed in practice, using data from two consumption surveys in Tanzania. Section 5 briefly concludes.

2. A BRIEF REVIEW OF THE LITERATURE

There are many sources of measurement error in expenditure surveys (Biemer et al., 1991). For example, recall errors occur when the information required is not easily retrieved from the respondent's memory and the information provided is inaccurate. Sometimes, it can assume the form of 'telescoping', i.e. the tendency to incorrectly perceive the temporal displacement of expenses incurred during the period under analysis. The reported answers can be also affected by the questionnaire (i.e. requiring information on a usual or a specific month or by the number of items considered in a list, and so on) and by the data-collection mode (i.e. the use of a diary, or the technology used for the interview, such as the Computer Assisted Personal Interview or the Computer Assisted Web Interview) and may determine significant variations in survey results (Tourangeau et al., 2000, Grosh and Glewwe, 2000, Friedman et al. 2016).

The impact of measurement errors on poverty and inequality measures has been studied less extensively. Many analyses have been conducted using a

[3] In regression analysis, it is well known that the presence of classical measurement errors in the explanatory variables leads to biased and inconsistent OLS estimators; in simple regression and correlation analysis, the bias assumes the form of attenuation bias, i.e. a tendency towards zero. In such cases, instrumental variables are a common tool for obtaining unbiased and consistent estimates (Chen et al., 2007).

case-by-case approach, by comparing survey data with administrative or other approximations of 'true data', and the conclusions cannot easily be extended to different contexts.

Cannari and D'Alessio (1993) and Gottschalk and Huynh (2010) evaluate the effects of measurement errors on the distribution of earnings and financial wealth respectively by comparing survey data with a benchmark containing 'true data' and find that survey data underestimate inequality. Similar results are obtained by D'Alessio and Neri (2015) who adopt a completely different approach, based on calibration techniques.

Cifaldi and Neri (2013), on analysing the reporting behaviour of Italian households, find that the misreporting of consumption has a different association with the reported amounts than with income: while under-reporting increases with declared income, there is no similar evidence for consumption. The explanation may be twofold: on the one hand, consumption is a less sensitive topic than income, because fiscal authorities are not interested in such amounts, on the other hand, consumption is more difficult to hide from an interviewer in a face-to-face interview.

These studies show that voluntary under-reporting, which is one of the main drivers of non-classical measurement errors, is usually much less significant for consumption than for income and wealth. This is the reason why in the following we will focus on classical measurement errors.

Widespread attention has been paid to the impact of outliers on poverty and inequality measures. Such studies have produced a much deeper knowledge of the sensitivity of various poverty and inequality measures to data contamination (Cowell and Flachaire, 2007). For example, these studies have made it clear that, generally speaking, inequality measures are more sensitive than poverty measures to extreme values. This is particularly true if poverty lines are exogenous (i.e. $1.25 per day) or are built on more stable in-sample statistics (i.e. median rather than mean) (Cowell and Victoria-Feser, 1996a; Cowell and Victoria-Feser, 1996b). Most of the time, the proposed estimators are obtained through the use of parametric models or by combining a parametric robust estimation of the upper tail of the distribution with the empirical data (the semi-parametric approach) (Victoria-Feser, 2000; Cowell and Victoria-Feser, 2007).

Measurement errors also affect the estimates of mobility in panel data (i.e. poverty dynamics). Either one looks at mobility tables describing the transitions from one state to another of a sample observed in two consecutive waves or one tries to estimate the growth of a variable observed against the initial value; (classical) measurement errors tend to overstate the actual mobility. Methods for obtaining mobility estimates accounting for the upward bias induced by measurement errors have been proposed by many authors (Luttmer, 2002; Neri, 2009; Glewwe, 2012; Burger et al., 2016; Lee et al., 2017).

Chesher and Schluter (2002) describe the approximated impacts on estimates of zero-mean measurement error, distributed independently of true income. They show that the Gini coefficient for an income distribution contaminated by measurement errors tends to be larger than the corresponding value obtained on the distribution of error-free. Measurement errors also raise the headcount poverty ratio, if the poverty line is below the mode of the distribution.

Lastly, it is worth noting some specific contributions dealing with the practice of estimating consumption items (i.e. yearly rents) by annualizing data collected over a short period of time (i.e. 1 month). Jolliffe and Serajuddin (2015) show that consumption data obtained by extrapolating to the whole year data collected on a weekly (or monthly) basis produce a higher headcount poverty ratio than that obtained using panel samples, for which yearly consumption data are derived by averaging the data obtained across multiple visits. Scott (1992) shows that this practice increases the variance of extrapolated annual expenditures, and proposes a method for adjusting such a bias. A similar adjustment is applied by Gibson et al. (2003) for the treatment of measurement errors on Chinese consumption data.

3. STATISTICAL TOOLS FOR ESTIMATING RELIABILITY

3.1 WHAT IS RELIABILITY?

Let X be a continuous variable measured with an error, so that the measure Y differs from the true value X by a random component: $Y = X + e$. If the disturbance has the following properties (called *homoscedastic, uncorrelated errors*):

$$E(e) = 0;\ E(X, e) = \sigma_{X,e} = 0;\ E(e, e) = \sigma^2_e$$

the variance of Y may be written as $\sigma^2_Y = \sigma^2_X \ / \ \lambda^2$ where $\lambda = \sigma_X / \sigma_Y$. The coefficient λ $(0 \leq \lambda \leq 1)$ is known as the *reliability index*; it expresses the share of variability in Y that belongs to the true phenomenon X (Lord and Novick, 1968), as opposed to the part due to the factors that contaminate the measurement process.[4]

[4] For a review of reliability analysis, see Webb et al. (2006).

The estimation of λ, as previously defined, would require knowledge of X, which can seldom be assumed. An easy way of obtaining an estimate of the reliability index is to resort to the test-retest, which involves a double measurement Y_1 and Y_2 of X in the same (approximate) survey conditions. Assuming homoscedastic and uncorrelated errors, the correlation coefficient between the two measurements Y_1 and Y_2 cannot be negative, and equals the square of the reliability index: $\rho_{y1,y2} = \lambda^2$.

In some cases it may be more appropriate to define a multiplicative measurement error, where the error can be assumed to be proportional to X, Y=X u, and u is a random variable independent from X, with E(u|x)=1 and variance σ^2_u. An interesting property of multiplicative errors is their ability to preserve the structural zeros in the distribution.

In such a case, one can rewrite the model as Y = X + X (u-1), and by posing w=X(u-1) one resorts to an additive error model Y = X + w, where E(w)=0, E(X, w) = $\sigma_{x,w}$ = 0 and $\sigma^2_w = (\sigma^2_x + \mu^2_x)\ \sigma^2_u$. As the error w in the additive formulation is uncorrelated with X (although with a heteroscedastic variance), the above equivalence between the correlation coefficient of two measures (Y_1 and Y_2) and λ^2 also applies.

It is worth noting that if the assumption E(e)=0 does not hold, where E(e)=δ, as may be the case in a particular survey design (i.e. the choice of the usual month consumption), the index λ only captures the variability of the two repeated measures Y_t and not their closeness to X (which is unknown). This means that the reliability index measured in this way evaluates the degree to which an instrument provides consistent measures; it does not indicate the instrument's truthfulness.

If we are dealing with categorical variables, the test-retest model needs to be revised (Biemer and Trewin, 1997). Let X be a categorical variable (with K categories) and Y its measurement. A reliability index for categorical features measured twice (Y_1 and Y_2) on the same set of n units is the fraction of units that are classified consistently: λ^*= tr (F)/n, where F is the cross tabulation of Y_1 and Y_2. Alternatively, one can resort to Cohen's kappa coefficient κ, which normalizes the share of observed matching cases with respect to their expected incidence if Y_1 and Y_2 are independent: $\kappa = (\lambda^* - \Sigma_i f_{i.} f_{.i}/n^2) / (1 - \Sigma_i f_{i.} f_{.i}/n^2)$.[5]

[5] Both the indices λ^* and κ can be adopted to assess the reliability of single categories of qualitative variables, computing them on the dummy variables by opposing each category to all the others (Biancotti et al., 2008).

3.2 RELIABILITY WITHIN A SINGLE SURVEY: DIARY DATA AS REPEATED MEASURES

The estimation of a reliability coefficient using data captured in a single survey is not an easy task. It may be unpleasant to ask a question more than once in the same survey and even if one does, it is likely that respondents tend to provide coherent answers, leading to an overestimation of reliability. In fact, the test-retest formula of λ relies on the assumption of uncorrelated errors; this assumption may be violated if the respondents realize that they have already answered the same question. Only a few studies provide reliability coefficients based on a double measurement in a single survey (e.g. Crossley and Kennedy, 2002).

In consumption surveys, however, the collection of data is often done by means of a diary, which can often be organized in a way that allows the computation of the reliability coefficient on repeated measures over time. For example, if the diary is kept by households over two weeks, it is possible to assume that, for the i-th generic household, the consumption of a good over a first period Y_1 (say those occurring during the first week, or on odd-numbered days) is a random variable with the same mean and variance as the consumption measured in a second period Y_2 (over the second week, or on even-numbered days); most of the time, the assumption of uncorrelated errors may reasonably hold.

In such a case, assuming homoscedastic and uncorrelated errors, one can simply estimate the reliability of the weekly consumption λ_1 as the square root of the correlation coefficient between the two measures. Moreover, according to the hypotheses described above, any other rearrangements of the daily measures into two halves may be used to compute the correlation coefficient (split-half approach) expression of reliability for weekly consumption.[6]

We have thus obtained an index measuring the reliability of data collected over half the period. The reliability over the whole period cannot be computed in the same way, as we have just a single measure. To obtain an estimate of the reliability λ_2 of the sum of the two weeks' consumption (Y_1+Y_2) starting from a measure of consistency between the two halves λ_1, we use the Spearman-Brown

[6] It is worth noting that the reliability coefficient measured in this way is not affected by a change over time in the average value of Y, as may happen in the case of a uniform fatigue effect across units. In such a case, the reliability index cannot account for the bias but still measures the variance across units correctly.

formula:[7] $\lambda_2 = [2\ \lambda_1]\ /\ [1 + \lambda_1]$. If you want to estimate the reliability λ_n referring to the sum of n weeks' consumption $(Y_1+Y_2+\ldots+Y_n)$, or more generally to the sum of n periods, the generalized formula applies:

According to the above formula, Table 1 shows how the reliability index modifies as the observation length widens. For example, if we estimate a reliability index of 0.6 for the weekly consumption, this corresponds to a reliability of 0.75 for the 2-week estimate and of 0.86 for the 4-week estimate[8]. As the weekly reliability increases, the gain obtained by extending the period for which the diary is kept reduces. This kind of information can help in evaluating the trade-off between a more stable estimation due to a longer diary data collection and the higher costs associated with such a choice.

It is important to bear in mind that the estimation of reliability as described above implies the independence of measurements over time (i.e. between the two periods considered). If this is not the case, as for example when the purchasing frequency is low and one purchase a day implies a reduced or even a zero value in the contiguous days, the reliability coefficients are underestimated by the exposed procedure, because of the presence of correlated errors.

Tab. 1: Reliability of repeated measures (Spearman-Brown formula)

Number of weeks (*)						
1	2	3	4	12	26	52
0.10	0.18	0.25	0.31	0.57	0.74	0.85
0.20	0.33	0.43	0.50	0.75	0.87	0.93
0.30	0.46	0.56	0.63	0.84	0.92	0.96
0.40	0.57	0.67	0.73	0.89	0.94	0.97
0.50	0.67	0.75	0.80	0.92	0.96	0.98
0.60	0.75	0.82	0.86	0.95	0.97	0.98
0.70	0.82	0.88	0.90	0.97	0.98	0.99
0.80	0.89	0.92	0.94	0.98	0.99	0.99
0.90	0.95	0.96	0.97	0.99	0.99	0.99
1.00	1.00	1.00	1.00	1.00	1.00	1.00

(*) The tables provides, for the various levels of reliability obtained for the weekly measure (column 1), the reliability expected if the data collection were extended over more weeks (according to the Spearman-Brown formula).

[7] The Spearman-Brown formula refers to the case in which the reliability of the average of n similar independent measures is derived from that of a single measure. See also Brown (1910) and Spearman (1910). Alternative estimators of reliability in the split-half scheme are found in Rulon (1939) and Guttman (1945).

[8] Of course, the reliability coefficients are in practice subject to a certain variability, and they may diverge from the Spearman-Brown formula as the classical hypotheses are not perfectly satisfied in the sample.

In general, the intertemporal variance of consumption survey data may be influenced by a purchasing frequency inadequate for the length of the observation window (see Gibson and Kim, 2011). For example, in a 2-week diary, households found with zero expenditure on clothes are presumably those who will have some clothing expenses during the rest of the year (and these amounts will probably be offset on average by other households for which positive expenses have been found during the observation window). If the window is too brief (compared with the purchasing frequency) the individual data will provide an unbiased estimate of the individual means but with a consistent standard error. Strictly speaking, this is not a matter of reliability, although it has similar effects on estimates.

Consumption surveys usually have a yearly reference period (S), while the estimates are often obtained on the basis of shorter periods of observation/recall (w), i.e. a 1- or 2-week diary or 1-3 months of information (i.e. for household bills), which are scaled up by means of an expansion factor (S/w). This practice does not take into account that while the average is not affected by the length of the collection period (w) over a homogeneous period, the variance of the reconstructed yearly consumption may depend on it (Deaton and Grosh, 1997).[9] With regard to this, the practice of spreading the full sample over the entire year with independent subsamples may be useful in the estimation of the mean (if it cannot be assumed to be constant over time) but does not help in any way in the estimation of the variance (or of the poverty and inequality indices).

Let us consider the case of the consumption of a good observed over a semester Y_1, whose amount is doubled to obtain a yearly estimate $Y'=2*Y_1$. In such a case, the variance of this estimator Y' is simply $Var(Y') = 4\ \sigma_1^2$. If we observe both the semesters and derive the yearly estimate by summing up these two components, $Y''= Y_1 + Y_2$, under the hypothesis of equal variance of the two semesters (i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$), we can derive the variance of Y'' as: $Var(Y'') = 2\ \sigma^2\ (1+\rho)$.

By comparing the two expressions we have that $Var(Y') \geq Var(Y'')$, where the equality only holds if the two components are perfectly correlated (i.e. $\rho=1$). Given the constraint $\sigma_1^2=\sigma_2^2=\sigma^2$, a perfect correlation implies the equality of two components $Y_1 = Y_2$. In other words, the variance of the extrapolated estimate Y' is always greater than that which would be obtained by collecting

[9] Clarke et al. (2008) show how an optimal trade-off between the higher precision characterizing short periods of recall and the greater stability affecting a wider range of observation can be determined.

data over the whole period, and will only be equal if there is no variation over time (i.e. $Y_1 = Y_2$) (Gibson et al., 2003; Gibson 2016).[10] The above result obviously holds even in the case of monthly (or weekly) estimates.

The conclusion is important for poverty and inequality analysis: if we can assume stability over time in the variance of a phenomenon, an assumption that is reasonable most of the time, then all other things being equal, inequality and poverty measures tend to have an upward bias both when the reliability of the measures is not perfect and when we use reduced observation/recall windows, given an intertemporal variability.[11] [12] It is particularly important that even if the variations over time are not due to measurement errors, they produce the same effect when the extrapolation strategy is applied, thus inflating the variance (and the poverty rates and inequality measures) of the phenomenon.

A corollary of the above statement is that, except for the unrealistic case of completely reliable and stable variables, inequality or poverty indices derived from consumption surveys with different observed periods are not immediately comparable, and would require some adjustment to take into account the abovementioned bias.

3.3 RELIABILITY WITHIN A SINGLE SURVEY: INTERNAL CONSISTENCY

Apart from diary data, it is rare to find repeated measures in the same consumption survey. Nonetheless, it is sometimes possible to assume that a set of variables is the expression of a unique latent variable. One could assume, for example, that the components of household consumption describe behaviour, which should have an internal coherence; marked deviations from this scheme could be an indicator of potential problems in the data.

The most widely used index for estimating reliability in terms of internal consistency among multiple items is Cronbach's alpha coefficient (Cronbach, 1951 and Cronbach and Shavelson, 2004). This coefficient that, contrary to what has been done so far, describes reliability in terms of variance rather than

[10] On the relationship between poverty measurement and the variability of economic outcomes within a year, see also Jolliffe and Serajuddin (2015).

[11] Following Istat (2016), which describes in detail the new methodology employed in the Italian Household Budget Survey and measures the impact of the changes on the estimates, the widening of the reference period for the consumption data collected in diaries has contributed significantly to lowering the relative poverty ratio.

[12] As we have already said, we have not considered other possible effects on estimates attributable to the length of the time period, such as the decrease in reporting due to a fatigue effect that affects average values.

in terms of standard deviation, can be written as: alpha = k r / (1 + (k-1) r), where r is the average of the k(k-1)/2 non-redundant correlation coefficients among the k items.[13] The alpha coefficient is also equal to the average of all possible split-half reliability estimates.

Under this framework, some descriptive statistics can be useful for deriving information on the reliability of single item variables. In the following, we will discuss some indicators, such as the correlation coefficients between each consumption component and the sum of all the other components, or the correlation coefficient between each consumption components and the values predicted by all the other components.

3.4 RELIABILITY OF PANEL DATA

In panel surveys, households are generally interviewed with a sufficient time lag to avoid any contamination of the first interview on the subsequent answers; for all the variables common to the waves, for which no changes may reasonably have occurred from one wave to another (i.e. time-invariant), a quantification of measurement error can be obtained by applying the test-retest formula.[14]

For time-varying variables, which are the majority of variables collected in consumption surveys, the analysis of measurement errors requires more sophisticated instruments because the quantities vary with time, and it is necessary to define models that distinguish actual change from movements induced by wrong measurements.

A method for estimating reliability indices using longitudinal data is provided by the simplex model (Heise, 1969; Alwin, 2007), within the more general framework of Structural Equation Modelling (SEM). Structural Equation Modelling (SEM) is a methodology for analysing relationships among variables widely used in social science research (Kaplan, 2000). It combines elements of other well-known statistical techniques, such as regression, path analysis and factor analysis, and is often used to model measurement errors, as it is able to study the links among measured variables and latent constructs.

[13] If the measures do not share the same mean and variance (i.e. measures are congeneric) the Cronbach coefficient is just a lower bound estimate of the true reliability.

[14] Biancotti et al. (2008) use Italian data to estimate the reliability of the variable measuring the floor area of residential dwellings, having selected the subsample of those households who did not move or incur extraordinary renovation expenses between the two survey waves. The reliability coefficient is $\lambda=0.80$.

In the simplex model, the reliability of data on time-varying quantities can be assessed, provided that at least three separate measurements of the variable on the same panel units are available; the separation of real dynamics from measurement error is obtained under mild regularity conditions (Biemer et al., 2009). This method (Heise, 1969), hypothesizes that the 3 variables X_1, X_2 and X_3 are measured by Y_1, Y_2, and Y_3 respectively, $Y_t = X_t + e_t$ (t=1,…3), with a homoscedastic, uncorrelated error e_t.

X_1, X_2 and X_3 are assumed to be pairwise related through independent, first-order autoregressive models, which do not need to be stationary:

$$X_1 = \delta_1 \,; \qquad X_2 = \beta_{21} X_1 + \delta_2 \,; \qquad X_3 = \beta_{32} X_2 + \delta_3$$

where $\beta_{t+1,t}$ is the autoregressive coefficient and δ_t (t=1,…3) is the process innovation. Innovations are uncorrelated pairwise.

Assuming constant reliability across the measures, the correlation coefficient between the observed values Y_t and Y_{t+1} can be written as $\rho_{Yt,Yt+1} = \lambda_Y^{\,2} \rho_{Xt,Xt+1}$, i.e. the correlation between X_t and X_{t+1} is attenuated by measurement errors both on Y_t and Y_{t+1}.

In such a case, the estimation of λ - assumed to be constant over the 3 waves - is obtained by means of the ratio of simple correlation coefficients:[15]

$$\lambda_Y = \sqrt{\frac{\rho_{Yt-1,Yt}\,\rho_{Yt,Yt+1}}{\rho_{Yt-1,Yt+1}}}$$

Under a first-order autoregressive assumption AR1, the above ratio should be equal to one if the variables are perfectly measured; when measurement errors are present, the ratio tends to decrease correspondingly.[16]

It is worth noting that the two parameters of the autoregressive model β_{21} and β_{32} do not need to be equal and may vary from one change to the next. What is supposed to be constant is the amount of measurement errors, an assumption that may reasonably be made in surveys conducted on a regular basis, with unchanged collection procedures.

[15] In the example provided by Biemer et al. (2009), the Heise measure is approximately the average of the measures obtained over the single waves with the alternative stationarity assumptions needed to identify the model.

[16] As observed by Biancotti et al. (2008), the Heise index measured under the AR1 hypothesis tends to be a downward-biased estimate of the reliability value if data follow an AR2 process.

4. MEASUREMENT ERRORS, POVERTY AND INEQUALITY

Having an estimate of the magnitude of measurement errors does not tell us which part of the variability we have to discard and how we can derive adjusted estimates.

A simple solution can be obtained by means of the method proposed by Scott (1992), which defines a transformation of collected data on a sample of n units C_i (i=1,...n) so that – by preserving the mean M - the standard deviation of the new variable C'_i is 0.9 times that of the old variable, $C'_i=M+(C_i - M)*0.9$. This transformation, which implies greater corrections in the tails of the distribution and lower corrections for values near to the mean, can help in understanding the possible impact of measurement errors on poverty and inequality estimates. The poverty and inequality measures computed on C' are usually lower than the corresponding measures computed on C.

We can also extend Scott's approach by adopting a different transformation of the data, such that only the residuals of a model are compressed in order to obtain the desired variability. For example, one could estimate the expected value of household consumption βX_i for a given set of known characteristics X_i, and then calibrate the variability of the residuals $e'_i = k\ e_i = k\ (C_i - \beta X_i)$, $(0<k<1)$ so that the standard deviation of the variable $C'_i= \beta X_i+e'_i$ is coherent with the estimated reliability λ. The value of k is obtained as $k = [(\sigma_T\ \lambda) - \sigma_E]/\ \sigma_R$, where σ_T, σ_E and σ_R are the standard deviation of the variables C_i, βX_i and $(C_i- \beta X_i)$ respectively. In this case too, the effects on poverty and inequality measures should predominantly be a reduction.

A different solution, which can be adopted when multiple items are available, is based on the use of Principal Component Analysis (PCA), a tool widely employed in denoising data. Following the singular value decomposition (Eckart and Young, 1936), we know that every matrix X with n observations and p variables (with n>p) can be fully decomposed based on the eigenvalues φ_m and eigenvectors u_m and v_m of the corresponding quadratic forms X'X and XX' respectively (with common non-zero eigenvalues φ_m but different eigenvectors u_m and v_m):

$$X = \Sigma_m\ \varphi_m\ u_m\ v_m' \qquad (m=1,\dots,p)$$

Keeping just the first k principal components (i.e. those corresponding to the highest eigenvalues) leads to a decomposition of X into one matrix of signal (X*) and one of noise (E), whose information can be discarded:

$$X = \Sigma_j\ \varphi_j\ u_j\ v_j' + E = X^* + E \qquad (j=1,\dots, k<p)$$

The first k principal components are the linear combinations of the original variables maximizing the variance, under the orthogonality constraint; they account for the maximum share of the global variance, expressed by the ratio $\tau=\Sigma_j \varphi_j /\Sigma_m \mathrm{Var}(x_m)$. However, τ is an average measure, as not all the variables are approximated in the same way. In this framework, the ratio of the standard deviation of each variable as approximated (by means of a linear prediction) by the first k principal components to its original standard deviation may be used as a measure of reliability. In geometrical terms, the reliability of each variable can be seen as the ratio of the length of the vector projected onto the optimal (in terms of explained variance) subspace of the first k principal components and the length of the same vector in the full p-dimensional space.

The choice of the number of principal components to retain is crucial. Sometimes, the 'eigenvalue one' rule of thumb is applied, which implies the retention of all the principal components whose variance is higher than that of the original (standardized) variables. In other cases, an analysis of the plot of the eigenvalues can help, for example, when it shows a clear drop in the explanatory power of the principal components. In some cases, information on the possible magnitude attributable to noise obtained by means of methods such as those illustrated in this paragraph can help with this task. From a practical point of view, as there is not always a unique and clear solution, a sensitivity analysis with various numbers of principal components is advised.[17]

A different solution is based on the Simulation-Extrapolation (SIMEX) method proposed by Cook and Stefanski (1994) for the estimation of regression parameters, which can easily be extended to poverty and inequality measures.

The method has two steps. In the first step, the method estimates the coefficient of interest (i.e. poverty ratios, Gini indices) on simulated data obtained by adding further measurement error to the available data. In this step, the researcher simulates various amounts of measurement errors many times (and if necessary, also different types of errors, such as additive or multiplicative).

Once the estimates of the poverty and inequality indices on contaminated data have been obtained, one can proceed with the second step, by estimating the relationship between measurement errors and the indices of interest. The adjusted estimates are obtained by extrapolating the trend back to the case of no measurement error.

The SIMEX method usually includes a graphical representation of the relationship between measurement errors and estimates that is able to account for such an adjustment.

5. AN EXAMPLE USING CONSUMPTION DATA FOR TANZANIA

5.1 THE DATA

In order to show an example of the methods described above, two main data sources have been considered in the paper: the Tanzania National Household Budget Survey (TNHBS) and the Tanzania National Panel Survey (TNPS). These surveys provide us with a complete set of data, able to show the potential of the tools described above.[18]

As regards the TNHBS, in the paper we use the 28-day diary data from the 2011-2012 wave, conducted on a sample of 10,186 households with completed interviews drawn from the 2002 Population and Housing Census frame. A stratified multi-stage sample design was used for this survey. At the first stage, the primary sampling units (PSUs) selected 400 enumeration areas (EAs). At the second stage, the EAs had an average of 133 households each, (155 for rural EAs and 94 for urban EAs). As some households were observed for longer than a month, only information concerning the first 4 weeks was retained in the analysis. In the paper estimates, sampling weights are used.

The Tanzania National Panel Survey (TNPS) is a survey conducted on a regular basis by the National Bureau of Statistics and the Ministry of Finance. The original sample, designed to be representative of the national, urban/rural, and main agro-ecological zones, consisted of about 3,200 households in the first 2008-2009 wave. The sample households were clustered in 409 EAs across Tanzania and Zanzibar.

In the second wave (2010-2011), the sample included the originally sampled households plus split-off households, while in the third wave (2012-2013) all the households interviewed during the previous two waves were contacted for the interview. Thus, the total sample of the last two waves is greater than that of the first wave (almost 4,000 units).

As the purpose of our analysis is to estimate the reliability of consumption measures, we have built our models only considering the approximately 1,000 households who did not change their composition across the 3 waves. In this way, the models accounting for changes over time can remain simple and deviations from the model can be attributed to measurement errors.

[17] For a discussion on this topic, see Gavish and Donoho (2014).

[18] Information on the TNHBS can be found here: https://www.nbs.go.tz/tnada/index.php/catalog/24 while on the TNPS are here: https://microdata.worldbank.org/index.php/catalog/2862.

Data refer both to nominal consumption, i.e. the total value of goods and services used by the respondents, and to real consumption, i.e. the nominal amount adjusted for temporal and spatial price deflators.

The attrition rate between the 2010/2011 and 2012/2013 waves was quite low, at around 3.5 per cent for households and 7.5 per cent for individuals.

5.2 RELIABILITY OF DIARY DATA (TNHBS)

In order to assess the reliability of the diary data collected by the TNHBS, we have grouped household expenses according to the week in which they occurred (1 to 4) and to the COICOP (Classification Of Individual COnsumption by Purpose) codes.

For every group of goods and services, the reliability index based on the correlation of weekly and bi-weekly household expenses has been computed (Table 2). As the diaries include 4 weeks, the averages of the 6 weekly and the 3 bi-weekly indices have been computed in order to summarize the results. Moreover, following the Spearman-Brown formula described above, the estimated reliability of the 4-week amount is presented.

In the data analysis, it is important to take into account that diary data are only a part of household consumption/expenditure, and that the share accounted for by the diary data may vary with the type of goods and services considered. For example, while food consumption items are fully noted in the diary, housing diary expenditures do not include the monthly (actual or imputed) rents for the house of residence and for other houses held as well as many other housing expenses collected by the questionnaire with reference to the last month (expenses for electricity, water and sewage services, waste collection and so on) or to the last three months (gas cylinders, charcoal, kerosene, coal and firewood). Analogously, fixed and mobile telephone bills and Internet subscriptions are not included in the communication expenses of the diary nor is the TV licence included in the recreation and culture expenses.

Conscious of these limitations, in the following we will discuss the reliability of the diary data alone, which do not fully represent the entire category except for food and beverages expenses.

Food and non-alcoholic beverages have an average weekly reliability of 0.7 that, according to the Spearman-Brown formula implies an estimated reliability for the corresponding 4-week diary figures of around 0.9. Both transport and communication have a similar 4-week reliability, of around 0.9, while alcoholic beverages show a reliability of around 0.86 and clothing and footwear a reliability of around 0.8. All the other figures are lower, in some cases as a clear effect of a typically low purchasing frequency (i.e. furnishings). For these latter items, the estimates of reliability – intended as the closeness of collected data to real values - are likely to be biased downwards; nonetheless, the low correlations signal instability over time, which may add undue variance to consumption estimates.[19]

It is worth noting that even if most of the expense items are complemented with other components from outside the diary, the limited reliability of some shares implies that additional variance is added to final estimates. Moreover, the collection of data for components outside the diary, such as for example the expenses for electricity, may also add variance to the total expenditure estimate, as they are collected on a last-month (or 3-month) basis and expanded to the year without taking measurement errors into account.

Values for the average bi-weekly indices that are consistently higher than the corresponding weekly measures signal a tendency to obtain more stable estimates as the diary period is extended. The reliability of total expenditures is equal to 0.729 for one week and 0.819 for two weeks; according to the Spearman-Brown formula, we derive an estimate of the reliability of total expenditures collected over four weeks of about 0.9.

If we look at data collected using diaries as panel data, we can also estimate reliability indices following the Heise model, allowing for some true variation over time on the basis of an AR1 model. The estimates obtained on the basis of the correlations observed between the expenditures over both the first 3 weeks and those of the last 3 weeks tend to agree, although with some exceptions (Table 2).

[19] The computation of average correlations is based on the assumption of equal reliability of weekly expenses. In our data, some descriptive analyses seem to suggest that this might not be entirely the case. For example, in 7 out of 11 types of goods and services considered, the correlation coefficients between the weekly expenses tend to increase, moving from the first to the second week and decreasing thereafter. The deviations are often not so important as to seriously affect our discussion based on an average measure; however, they could reflect both an initial learning effect in compiling the diary and a subsequent fatigue effect.

Tab. 2: Reliability of diary aggregates

Consumption aggregates	Reliability based on the average weekly correlation	Reliability based on the average bi-weekly correlation	Estimate of reliability of 4 weeks' expenditure *	Estimate of reliability of 4 weeks' expenditure **	Heise model weekly reliability coefficients – (weeks 1 to 3)	Heise model weekly reliability coefficients – (weeks 2 to 4)
Food and non-alcoholic beverages	0.707	0.784	0.906	0.879	0.867	0.812
Alcoholic beverages	0.612	0.751	0.863	0.858	0.698	0.585
Clothing and footwear	0.540	0.677	0.825	0.808	0.708	0.478
Housing, water, electricity, gas	0.486	0.608	0.791	0.756	0.874	0.760
Furnishings	0.279	0.382	0.608	0.553	0.214	0.304
Health	0.316	0.428	0.649	0.599	0.367	0.590
Transport	0.693	0.799	0.900	0.888	0.798	0.758
Communication	0.675	0.789	0.892	0.882	0.705	0.673
Recreation and culture	0.221	0.286	0.532	0.445	0.514	0.343
Education	0.063	0.077	0.213	0.144	0.383	0.015
Other goods and services	0.483	0.617	0.789	0.763	0.423	0.434
Total expenditures	0.729	0.819	0.915	0.901	0.896	0.828

* Obtained by applying the Spearman-Brown formula shown in the text to the reliability based on the average weekly correlation. ** Obtained by applying the Spearman-Brown formula shown in the text to the reliability based on the average bi-weekly correlation.

5.3 RELIABILITY OF PANEL DATA (TNPS)

Table 3 shows the reliability coefficients computed in different ways for twelve main components of total household consumption and equivalent consumption[20] collected in the TNPS.

The first column refers to the coefficient obtained by applying the Heise model to household consumption components. As these estimates, as well as

[20] The equivalent household consumption is obtained by dividing the household consumption by a coefficient (equivalence scale) which makes it possible to take into account the effect of economies of scale. Equivalent consumption allows the comparison of the corresponding welfare across households of different sizes and compositions.

Tab. 3: Reliability coefficients for some expenditure aggregates

Consumption aggregates (a)	Heise model coefficients - Household consumption		Heise model coefficients - Equivalent household consumption	
	Value	Rankings	Value	Rankings
1. Food and non-alcoholic beverages: at home and away from home	0.780	0.810	0.820	0.757
2. Alcohol and tobacco: at home and away from home	0.764	0.787	0.812	0.793
3. Food, beverages, alcohol and tobacco: at home	0.719	0.813	0.650	0.695
4. Food, beverages, alcohol and tobacco: away from home	0.877	0.656	0.747	0.661
5. Utilities: water, kerosene, lighting	0.935	0.893	0.905	0.905
6. Furnishings and household expenses	0.870	0.671	0.756	0.666
7. Health	0.854	0.576	0.461	0.532
8. Transportation	0.622	0.665	0.531	0.660
9. Communications	0.762	0.876	0.714	0.880
10. Recreation	0.232	0.318	0.378	0.319
11. Education	0.996	0.968	0.793	0.968
12. Other consumption	0.654	0.833	0.748	0.847
13. Total consumption - nominal	0.905	0.882	0.919	0.842
14. Total consumption – real	0.884	0.867	0.899	0.826

(a) Aggregates 1+2 = 3+4 = Total food consumption.

the other estimates considered in the table, are computed by evaluating the heterogeneity of the answers provided by panel households over time, only the approximately 1,000 households who did not change their composition were considered. As a robustness check, the second column shows the reliability coefficients computed on the rankings, which are less influenced by outliers.

The results show quite a satisfactory reliability of total consumption, with estimates of just below 0.9 both for nominal and real figures. In other words, 90 per cent of the variability of these indicators is consistent between the measures while the remaining part is attributable to measurement error.

However, reliability is not constant across the consumption components. It is higher for both utilities and education, which account for a few expenses on a more regular basis. On the other hand, the lowest reliability is observed for recreational consumption, which is more difficult to capture due to its lower regularity over time and higher granularity across household members. Modest reliability also characterizes transportation and (in most estimates) health consumption.

The reliability of food consumption consumed both at home and away from home (excluding alcoholic beverages), is around 0.8, quite similar to the estimates of the previous paragraph.

The reliability of total consumption does not improve when the least reliable variables are excluded from the sum. For example, the sum of all consumption items excluding recreation provides an aggregate whose reliability is just a little lower than that of the complete aggregate; also excluding transportation or health consumption from the total slightly decreases reliability. In general, adding up items improves the reliability of aggregates.

As already observed for health, the estimates of reliability coefficients do not always display stable behaviour. Heise coefficients computed on consumption values and on rankings only show a moderate agreement; some differences between these estimates are quite large (e.g. 0.877 and 0.656 for food consumption away from home, or 0.854 and 0.576 for health consumption). On the whole, the analysis of the different coefficients does not always provide a clear picture or eliminate any doubts as to the real situation.

A greater instability characterizes the reliability coefficients computed on more detailed food consumption items (Table 1A in Appendix A reports some examples). In fact, the correlation coefficients of specific food consumption between two consecutive waves – always computed only on households who did not change their composition - are often quite low, around 0.2 on average and only in a few cases are they significantly higher (never greater than 0.625).[21]

By collecting specific consumption items over a single week, the survey captures the behaviour of households that is only weakly confirmed in the second wave. As we discussed earlier, although the short reference period reduces the memory biases and other forms of contamination, and they are presumably close to the actual data, the collected data do not provide an accurate picture of the consumption behaviour of that household over the entire year. In any case, low correlations over the waves imply a greater instability in the estimates of Heise coefficients, which may even go outside the range 0-1 (as happens in almost a quarter of the cases).

[21] As a comparison, the average of one-lag correlation coefficients computed on the main consumption aggregates is around 0.55.

As we have already said, a different approach to dealing with the reliability of answers relies on the analysis of internal consistency. In consumption surveys, several instruments developed for this kind of data can be fruitfully used.

Table 4 shows several indices that can help in understanding the reliability of the collected data. The first column shows the correlation between the values of the specific variable in the row and the sum of all the other components of total household consumption. The higher the value, the more the data contained in the variable is coherent with the sum of all the other components. The second column shows the correlation of the component in the row and the predicted values of the multiple regression with all the other components. As the least squares solution is the linear combination maximizing the predictability of the dependent component, this measure is always higher than that computed on the sum of the components. The third and the fourth columns of the table show the share of the standard deviation of the component that is accounted for by the first and the first three principal components respectively. As the first principal component is a linear combination of all the consumption items, including that on the row, it tends to be higher than the previous two measures, although this is not always the case (for example, see recreation consumption). The same four measures are then computed for the equivalent consumption.

On the whole, the picture drawn by these indicators is coherent with that described above, mainly when considering the Heise indices computed on the rankings rather than on the values.

Recreation and health consumption show low indices of internal consistency, confirming the results obtained with the Heise model. A low internal consistency index also characterizes ‘Alcohol and tobacco: at home and away from home’ which, on the contrary, showed a good performance in terms of coherence over time. On the other hand, a good performance is found for utilities, communication, other consumption and food (excluding alcohol and tobacco), largely confirming the previous results.

In the comparative analysis of these results, it is worth taking into account that the reliability measures have been computed in the two frameworks under different assumptions. Clearly, random errors affecting indicators imply both a reduced ability of an AR1 model to account for data and lower internal consistency of data. A different performance of an indicator under the two frameworks could signal some deviation from the hypotheses on which the models are built.

Tab. 4: Internal consistency of 2012 consumption aggregates

Consumption aggregates (a)	Household consumption				Equivalent consumption			
	Correlation with the sum of all the other components	Multiple correlation with all other components	Correlation with the first principal component	Multiple correlation with the first three principal components	Correlation with the sum of all the other components	Multiple correlation with all other components	Correlation with the first principal component	Multiple correlation with the first three principal components
1. Food and non-alcoholic beverages: at home and away from home	0.630	0.648	0.833	0.923	0.613	0.681	0.842	0.909
2. Alcohol and tobacco: at home and away from home	0.234	0.238	0.323	0.832	0.205	0.196	0.282	0.744
3. Food, beverages, alcohol and tobacco: at home	0.410	0.517	0.663	0.832	0.235	0.529	0.560	0.909
4. Food, beverages, alcohol and tobacco: away from home	0.383	0.498	0.623	0.813	0.291	0.495	0.614	0.934
5. Utilities: water, kerosene, lighting	0.631	0.723	0.760	0.796	0.648	0.732	0.772	0.826
6. Furnishings and household expenses	0.489	0.553	0.613	0.682	0.484	0.533	0.604	0.647
7. Health	0.296	0.316	0.357	0.552	0.237	0.398	0.349	0.668
8. Transportation	0.549	0.617	0.682	0.700	0.517	0.609	0.659	0.679
9. Communications	0.639	0.694	0.757	0.767	0.681	0.710	0.773	0.776
10. Recreation	0.190	0.293	0.282	0.756	0.195	0.266	0.268	0.485
11. Education	0.517	0.568	0.633	0.657	0.274	0.329	0.370	0.516
12. Other consumption	0.679	0.747	0.794	0.814	0.676	0.725	0.771	0.805

(a) Aggregates 1+2 = 3+4 = Total food consumption. The two pairs of food aggregates have been used alternately in the computation of all the indices of the table.

This could be the case of 'alcohol and tobacco', for which the indices based on the models give a picture of satisfying reliability, which is not confirmed when looking at the coherence with other consumption items. A similar result is found for health consumption. This suggests that these kinds of consumption do not share the same latent variable, as the conditions at their root (the need to smoke or drink, or poor health conditions) are only partially related to the consumption behaviour.

Furthermore, the reliability indicators for food items based on internal consistency largely confirm the results obtained with the above models (Table 2A in Appendix A).

5.4 ADJUSTING POVERTY AND INEQUALITY MEASURES IN THE 2012 TNPS

By adopting a standard definition of the relative poverty rate, i.e. the count of households whose equivalent nominal consumption falls below half the median, we find a share of 15.8 per cent of households in the 2012 TNPS. As we have shown in Table 3, the total consumption collected in the TNPS has a reliability of around 0.9, that is to say that 10 per cent of the standard deviation is due to measurement errors and should be accounted for.

We first apply the method proposed by Scott (1992), which defines a transformation of collected data Y_i so that – preserving the mean - the standard deviation of the new variable X_i is 0.9 times that of the old variable. The poverty rate obtained from these transformed data is much lower than that obtained with the original data (8.2 per cent).[22] The Gini concentration index would also be greatly reduced, from 0.436 to 0.393 (Table 5).

Tab. 5: Poverty ratios and Gini indices for real and simulated equivalized household consumption distributions*

Variable	Share of households below the poverty threshold	Gini index
Observed equivalent consumption	15.8	0.436
Scott's adjustment	8.2	0.393
Modified Scott's adjustment	13.1	0.393
1 principal component	10.4	0.417
2 principal components	9.8	0.419
SIMEX (additive error)	7.8	0.334
SIMEX (multiplicative error)	13.7	0.415

* Data from the TNHBS, 2012.

[22] In general terms, the headcount poverty ratio defined in absolute terms can be even more affected by such a transformation, because the poverty line does not shift with the distribution. The impact depends on the mass of the distribution around the poverty line.

We also apply the extension of Scott's approach as illustrated in section 4, consisting in a transformation of the data such that only the unexplained variance of a model is compressed in order to obtain the desired variability. In our case, we regress the 2008 and 2010 data on the 2012 expenditures and use the predicted values instead of the unconditional mean in the above formula; the coefficient for the compression of the residuals is adjusted accordingly. In such a case, the poverty rates for the 1,000 homogeneous panel sample units decline by 2.7 percentage points. The change from original to adjusted estimates is smaller than that obtained in the previous adjustment but is still considerable. The Gini concentration index is reduced to 0.393, as in the previous experiment.

A further experiment consisted in finding an approximation of the components of household equivalent expenditures by means of a Principal Component Analysis. Table 6 shows, for various possible approximations (with 1, 2, …, k principal components), the share of the standard deviation accounted for by the principal components for each variable, which provides information about the reliability of each item. It also shows the poverty rate obtained on the total equivalent expenditures derived by adding up the values predicted by the k principal components.

The poverty rate is around 10 per cent when the first principal components are considered (up to 8); it grows to 15.8 per cent when all the possible components are considered (i.e. no errors are considered). As a possible criterion for selecting the number of principal components to retain, we observe that with the first 2 principal components we obtain an approximation of the original components of the total expenditures that, with some exceptions (i.e. education), is quite close to that derived using the Heise model (Figure 1). In other words, the two methods seem to converge independently towards similar results. As in the previous examples, there are clear indicators that poverty rates could be significantly overestimated when using standard estimators. As to the Gini concentration index, when considering up to 5 principal components it is always around 0.418, higher than in the previous two experiments but lower than the original value (0.436).

Although the only purpose of the experiments is to provide an indication of the possible impacts of measurement errors in poverty and inequality measures, the results converge towards the conclusion that the estimates that do not take measurement errors into account can be upward-biased.

Tab. 6: Principal Component Analysis of 2012 main expenditmure items

	Share of the standard deviation accounted for by the first k principal components							
	1	2	3	4	5	6	...	10
Food and non-alcoholic beverages	83.3	87.3	91.0	91.1	91.1	98.0	...	100.0
Alcohol and tobacco	26.9	73.7	73.8	74.4	75.4	99.7	...	100.0
Utilities: water, kerosene, lighting	81.2	84.2	88.2	88.3	88.4	89.5	...	100.0
Furnishings and household expenses	60.2	63.1	63.4	67.2	67.3	99.4	...	100.0
Health	42.9	59.9	77.6	87.4	90.6	99.4	...	100.0
Transportation	64.8	64.8	66.0	66.3	68.8	92.9	...	100.0
Communications	79.1	79.6	79.6	79.6	81.3	86.0	...	100.0
Recreation	25.5	26.8	49.3	96.6	96.7	99.9	...	100.0
Education	35.6	42.7	47.7	52.7	98.7	99.4	...	100.0
Other consumption	81.6	86.0	87.0	89.5	89.8	90.1	...	100.0
Poverty rate *	10.4	9.8	9.7	9.3	9.5	10.4	...	15.8
Gini index **	0.417	0.419	0.419	0.417	0.418	0.422	...	0.436

* Household poverty rate computed on the total equivalent expenditures obtained as the sum of the estimated components predicted by the k principal components.

** Gini index computed on the total equivalent expenditures obtained as the sum of the estimated components predicted by the k principal components.

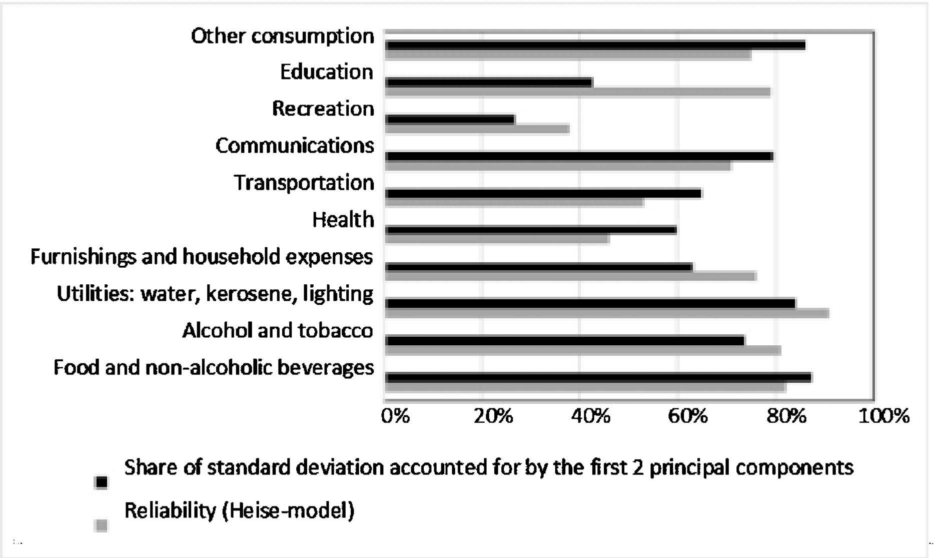


Fig. 1: Share of standard deviation accounted for by the first 2 principal components and the reliability coefficients obtained using the Heise model

Last, I apply the SIMEX approach to the total equivalized household consumption C collected in the 2012 TNPS.

In the simulation step, I hypothesize that the variable C is free of error and I consider 6 random variables $e_1,\ldots, e_6$, uncorrelated with C and with zero mean and standard deviations calibrated in order to simulate 6 measures $C^+_1=C+e_1,\ldots,C^+_6 = C+e_6$, with reliability indices equal to 0.95, 0.90, 0.85, 0.80, 0.75 and 0.70 respectively.

In order to account for multiplicative measurement errors, I also simulate 50 times 6 random variables $u_1,\ldots, u_6$, uncorrelated with C and with a mean equal to 1, such that the simulated measures $C^*_1=C_1u_1$, …, $C^*_6 = C_6u_6$, have reliability indices equal to 0.95, 0.90, 0.85, 0.80, 0.75 and 0.70 respectively.

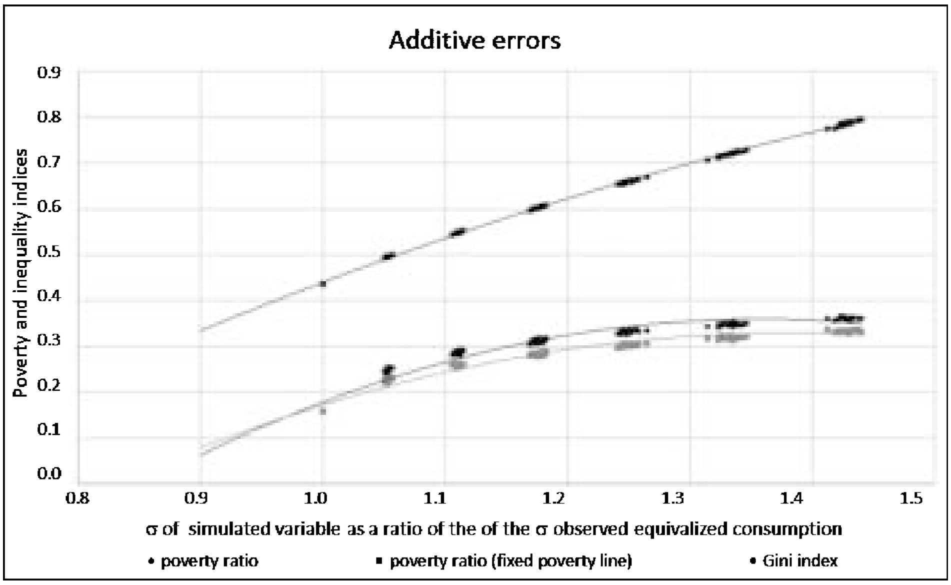
For every simulation, poverty and inequality measures are computed: in this way, the bias induced by the measurement errors can be analysed. The simulated values are shown in Figure 2, together with a line accounting for the average relationships between measurement errors and poverty and inequality measures.[23] Once the two lines have been obtained, one for the additive and the other for the multiplicative errors, we can derive the estimates corresponding to the standard deviation of the error free consumption (in our exercise, hypothesized as being equal to 0.9 of the measured variable C).

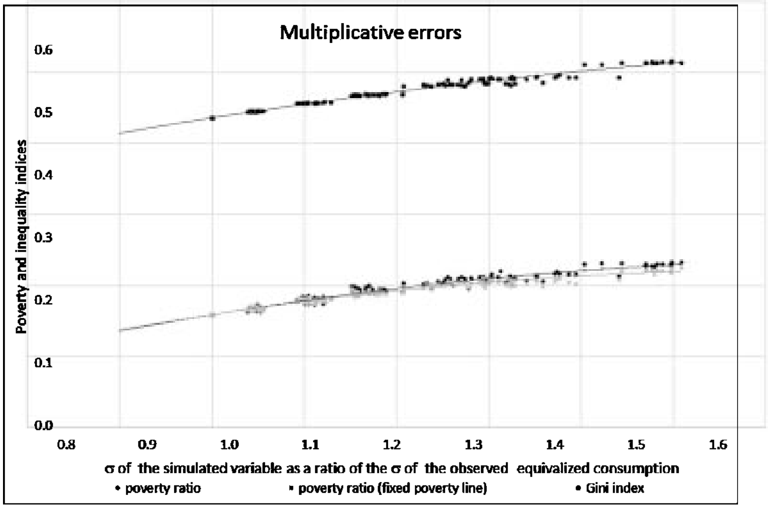
By definition, the variability of C^+_1, …, C^+_6 and C^*_1, …, C^*_6 is higher than that of C; so we should observe a growth in the Gini index as the reliability decreases and the distribution of measured consumption flattens.[24] The extrapolated indices at a value of the standard deviation equal to 90 per cent of that of the variable C (0.334 and 0.415 for additive and multiplicative errors respectively) are lower than those observed for C (0.436); the reduction is much more intense in for additive measurement errors than in case of multiplicative measurement errors.

The impact of measurement errors on poverty ratios is instead less trivial. The relative poverty rate is the share of households whose equivalent nominal consumption falls below half of the median; adding random noise to the distribution of C may also alter the median and the corresponding poverty threshold.

[23] A quadratic equation was used for extrapolating the simulated data.

[24] Simulated data with the additive model can be sometimes negative. In such a case, the Gini index can be greater than one and should just be interpreted as a relative measure of variability. However, the result of growing concentration with decreasing reliability is also confirmed when negative values are set at zero.





(*) Simulated data are obtained by adding random measurement errors to the observed equivalent consumption. The value of 1.1 on the x axis means that the standard deviation of simulated data is 10% greater than that of observed data, 1.2 means that the standard deviation is 20% greater, and so on. The poverty and inequality indices for value of 1 on the x axis are the estimates obtained on the observed values, while those corresponding to 0.9 are the extrapolated SIMEX estimates (by means of a quadratic function), under the hypothesis that 10% of standard deviation of the observed equivalized consumption is noise (reliability 0.9).

Fig. 2: Poverty and inequality measures for observed and simulated data of equivalent consumption (*)

For this reason, we first analyse the case of the fixed threshold, looking at the changes in the share of households whose simulated equivalized consumption $C^{+}i$ (i,1,…,6) and C*i (i,1,…,6) falls below the poverty threshold of C. As expected, with a fixed threshold, the share of poor households steadily increases as the reliability decreases. The extrapolated estimates for a standard deviation equal to 90 per cent of that of C are equal to 6.3 and 13.3 per cent for additive and multiplicative errors respectively. In this case too, the reduction is more intense for the additive model.

However, to have an idea of the full impact of measurement errors on poverty indices, we have to see how the median modifies on moving from C to $C^{+}i$ (i,1,…,6) and C*i (i,1,…,6).

In simulated data, the median rises steadily as reliability decreases. This result, which is not ensured for any data distribution, should hold when dealing with quite regular asymmetrical distributions with positive skewness, for which mode<median<average, as for example the log-normal distribution. Intuitively, it can be explained by observing that in such a case, the median lies in the descending part of the distribution, and adding symmetrical errors to observed values implies a greater probability of exceeding the median for observed data below the median than the opposite case. This explains why in simulated data the share of households below the poverty threshold rises even more markedly than observed above as the reliability decreases. The extrapolated estimates for the poverty ratio become 7.8 and 13.7 per cent for additive and multiplicative errors respectively.

The results obtained for both the poverty ratio and the Gini index with multiplicative errors in the SIMEX method can be considered more conservative, and may therefore be preferable.

In the end, the results provided by the different methods employed for the adjustments vary from one method to another or depending on the additive or multiplicative model considered. However, all the experiments carried out in this section clearly confirm that measurement errors in consumption data may significantly bias poverty and inequality measures upwards.

6. CONCLUSIONS

The paper analysed the measurement errors affecting the most important variables collected in two consumption surveys carried out in recent years in Tanzania: the Tanzania National Household Budget Survey (TNHBS) and the Tanzania National Panel Survey (TNPS). These surveys gave us the chance to study measurement errors using both diary and panel data, and to address general issues regarding the relationships between the quality of data and the estimation of poverty and inequality measures.

According to our estimates and models, all the variables collected in these surveys are affected by a share of measurement errors, not, homogeneous across the items, however, and that may be affected by the way the data are collected.

In the diary data collected in the TNHBS, food, transport and communication expenditures show quite good reliability (around 0.8); instead lower reliability characterizes furnishings and – for the share collected by diary – education expenses. The reliability of single food consumption items is instead generally lower, and presumably affected by the consumption/purchasing frequency.

For TNPS data, the reliability is quite high for total consumption (around 0.9); food and non-alcoholic beverages have reliability values of around 0.8; lower reliability is found for some components, such as expenditure on recreation, for which lower regularity over time and more expenses spread across individuals can be presumed. Low reliability is also found for some estimates concerning expenditure on health, mainly when the internal consistency approach is adopted. The result underlines that these expenses are not fully driven by the same latent variable of other expenditures.

Measurement errors tend to inflate the variance of collected variables. Moreover, the common practice of extrapolating data observed over a short period of time (i.e. one month) to the whole reference period (i.e. the year) can be seen as a further measurement error, inflating the variance of indicators. Other things being equal, the longer the length of the observation period (i.e. the period for which diary data are collected) the lower the variance of collected data.

An assessment of the measurement errors contained in the data is important both for data producers, who may find useful information for improving the data collection methods employed and for data analysts. The researchers who use consumption micro-data should properly take into account that – other things being equal - the higher the reliability the lower the inequality measures. Moreover, poverty indicators also tend to be biased upwards by the measurement errors.

In the paper, various methods for obtaining more robust estimates are provided. In particular, the method proposed by Scott, the Principal Component

Analysis and the SIMEX method have been described. The results obtained for Tanzania seem satisfying, but more research in this field is needed.

Given that a certain degree of measurement errors is unavoidable in sample surveys, all the above considerations suggest both the adoption of best practices in the collection of survey data, in order to improve the reliability of data, and a move towards a standardization of the collection methods employed, which reduces the risks of contaminating the comparisons with spurious effects. In particular, in sample surveys conducted on a regular basis, the improvements that usually occur in the data collection procedures could reduce measurement errors over time, and thus produce a bias in the trend of poverty and inequality measures. In such a case, it is important to adopt a strategy for allowing the measurement of the impact of the changes in the data collection methods on the estimates.

Appendix A – Reliability of food consumption items

Tab. 1A: Reliability coefficients of some food consumption items (Heise model), 2008-2013

Item Code	Food consumption item (selected items)	Yes/No	Values	Rankings
101	Rice (paddy)	0.544	a	a
102	Rice (husked)	0.665	0.820	0.796
103	Maize (green, cob)	0.716	0.308	0.269
104	Maize (grain)	a	a	0.468
105	Maize (flour)	0.622	0.627	0.714
106	Millet and sorghum (grain)	0.293	a	a
107	Millet and sorghum (flour)	0.379	0.325	0.356
108	Wheat, barley grain and other cereals	0.575	0.638	0.643
109	Bread	0.699	0.624	0.741
110	Buns, cakes and biscuits	0.763	0.564	0.723
111	Macaroni, spaghetti	0.393	0.297	0.403
112	Other cereal products	0.490	0.414	0.322
301	Sugar	0.752	0.767	0.794
302	Sweet potatoes	a	0.288	a
1001	Cooking oil	0.818	0.670	0.780
1002	Butter, margarine, …	0.471	0.412	0.498
1003	Salt	0.855	0.463	0.825
1004	Other spices	0.437	0.364	0.445
1101	Dry tea	0.802	0.599	0.779
1102	Coffee and cocoa	0.365	a	0.152
1104	Bottled/canned soft drinks	0.518	0.498	0.538
1107	Local brews	0.552	0.540	0.503
1108	Wine and spirits	a	a	a

(a) Coefficients outside the range (0-1).

Tab. 2A: Internal consistency of some 2012 food consumption items

Item Code	Food consumption item (selected items)	Correlation with the sum of all the other components	Multiple correlation with all other components	Correlation with the first principal component	Multiple correlation with the first three principal components
101	Rice (paddy)	0.003	0.161	0.012	0.012
102	Rice (husked)	0.581	0.695	0.536	0.536
103	Maize (green, cob)	0.162	0.298	0.169	0.169
104	Maize (grain)	0.035	0.215	0.056	0.056
105	Maize (flour)	0.187	0.373	0.185	0.185
106	Millet and sorghum (grain)	0.028	0.122	0.036	0.036
107	Millet and sorghum (flour)	0.121	0.259	0.143	0.143
108	Wheat, barley grain and other cereals	0.268	0.602	0.245	0.245
109	Bread	0.429	0.578	0.393	0.393
110	Buns, cakes and biscuits	0.284	0.436	0.236	0.236
111	Macaroni, spaghetti	0.291	0.469	0.312	0.312
112	Other cereal products	0.130	0.242	0.121	0.121
301	Sugar	0.468	0.694	0.448	0.448
302	Sweet potatoes	0.289	0.440	0.293	0.293
1001	Cooking oil	0.613	0.734	0.599	0.599
1002	Butter, margarine, …	0.306	0.484	0.356	0.356
1003	Salt	0.194	0.465	0.188	0.188
1004	Other spices	0.420	0.542	0.409	0.409
1101	Dry tea	0.476	0.605	0.426	0.426
1102	Coffee and cocoa	0.177	0.341	0.197	0.197
1104	Bottled/canned soft drinks	0.442	0.645	0.438	0.438
1107	Local brews	0.016	0.282	0.021	0.021
1108	Wine and spirits	0.040	0.223	0.044	0.044

REFERENCES

Alwin, D.F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement*. Wiley.

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*, Vol. 54, No. 3, pp. 466-493.

Biemer, P.P., Christ, S.L., Wiesen, C.A. (2009). A general approach for estimating scale score reliability for panel survey data. *Psychological Methods*, Vol 14, No. 4, pp. 400-412. http://dx.doi.org/10.1037/a0016618

Biemer, P.P., Groves, R.M., Lyberg, L., Mathiowetz, N.A., Sudman, S. (1991). *Measurement errors in surveys*. Wiley.

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

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, Vol. 3, No. 3, pp. 296–322.

Burger, R.P., Klasen, S., Zoch, A. (2016). Estimating income mobility when income is measured with error: the case of South Africa. REDI3x3 Working paper, No.14, April.

Cannari, L., D'Alessio, G. (1993). Non-reporting and Under-reporting Behavior in the Bank of Italy's Survey of Household Income and Wealth. In *Bulletin of the International Statistical Institute*, Vol. LV, No. 3, Firenze, pp. 395-412.

Chen, X., Hong, H., Nekipelov, D. (2007). *Measurement Error Models*. http://web.stanford.edu/~doubleh/eco273B/survey-jan27chenhandenis-07.pdf

Chesher, A., Schluter, C. (2002). Welfare Measurement and Measurement Error. *The Review of Economic Studies*, Vol. 69, No. 2, April, pp. 357-378.

Cifaldi, G, Neri, A. (2013). Asking income and consumption questions in the same survey: what are the risks? *Temi di Discussione (Working papers)*, No. 908, Banca d'Italia.

Clarke, P.M., Fiebig, D.G., Gerdtham, U.G. (2008). Optimal recall length in survey design, *Journal of Health Economics*. Vol. 27, No. 5, pp. 1275–1284.

Cook, J.R., Stefanski, L.A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* Vol. 89, No. 428, pp. 1314-1328.

Cowell, F.A., Flachaire, E. (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics*, Vol. 141, No. 2, pp. 1044–1072.

Cowell, F.A., Victoria-Feser, M.P. (1996a). Poverty measurement with contaminated data: a robust approach. *European Economic Review*, Vol. 40, No. 9, pp. 1761-1771.

Cowell, F.A., Victoria-Feser, M.P. (1996b). Robustness properties of inequality measures. *Econometrica*, Vol. 64, No. 1, pp. 77-101.

Cowell, F.A., Victoria-Feser, M.P. (2007). Robust stochastic dominance: A semi-parametric approach. *Journal of Economic Inequality*, Vol.5, No. 1, pp. 21-37.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*. Vol. 16, No. 3, pp. 297–334.

Cronbach, L.J., Shavelson, R.J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, Vol. 64, No. 3, pp. 391–418.

Crossley, T.F., Kennedy, S. (2002). The reliability of self-assessed health status. *Journal of Health Economics*, Vol. 21, No. 4, pp. 643-58.

D'Alessio, G., Neri, A. (2015). Income and wealth sample estimates consistent with macro aggregates: some experiments. *Questioni di Economia e Finanza (Occasional Papers)*, No. 272, Banca d'Italia.

Deaton, A., Grosh, M. (1997). *Designing Household Survey Questionnaires for Developing Countries: Lessons from 10 Years of the Living Standards Measurement Study*. Edited by M. Grosh and P. Glewwe, The World Bank.

Eckart, G., Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, Vol. 1, No. 3, pp. 211-218.

Friedman, J., Beegle, K., De Weerdt, J., Gibson, J. (2016). Decomposing Response Errors in Food Consumption Measurement: Implications for Survey Design from a Survey Experiment in Tanzania. *World Bank Policy Research Working Paper*, No.7646, April.

Gavish, M., Donoho, D.L. (2014). *The Optimal Hard Threshold for Singular Values is 4/sqrt (3)*. arXiv:1305.5870v3 [stat.ME], 4 June, http://arxiv.org/pdf/1305.5870.pdf

Gibson, J. (2016). Measuring Chronic Hunger from Diet Snapshots: Why 'Bottom up' Survey Counts and 'Top down' FAO Estimates Will Never Meet. *Working Papers in Economics*, No. 7, University of Waikato, Department of Economics.

Gibson, J., Kim, B. (2011). How reliable are household expenditures as a proxy for permanent income? Implications for the income–nutrition relationship. *Department of Economics Working Paper Series*, No. 3, Hamilton, New Zealand: University of Waikato.

Gibson, J., Huang, J., Rozelle, S. (2003). Improving Estimates of Inequality and Poverty from Urban China's Household Income and Expenditure Survey. *Review of Income and Wealth*, Vol. 49, No. 1, pp. 53-68.

Glewwe, P. (2012). How Much of Observed Economic Mobility is Measurement Error? IV Methods to Reduce Measurement Error Bias, with an Application to Vietnam. *World Bank Economic Review*, Vol. 26, No. 2, pp. 236-264.

Gottschalk, P., Huynh, M. (2010). Are Earnings Inequality and Mobility Overstated? The Impact of Non-Classical Measurement Error. *The Review of Economics and Statistics*, Vol. 92, No. 2, pp. 302-315.

Grosh, M., Glewwe, P. (2000). *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study*. Vol. 1, 2, and 3, Oxford University Press (for the World Bank).

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, Vol. 10, No. 4, pp. 255-282.

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

Heise, D. (1969). Separating Reliability and Stability in Test-Retest Correlation. *American Sociological Review*, Vol. 34, No. 1, pp. 93-101.

Istat (2016). *La nuova indagine sulle spese per consumi in Italia*. Edited by D. Grassi, N. Pannuzi and C. Freguja, Metodi – Letture Statistiche.

Jolliffe, D., Serajuddin, U. (2015). Estimating poverty with panel data, comparably: an example from Jordan. *World Bank Policy Research Working Paper*, No. 7373.

Kaplan, D. (2000), *Structural Equation Modeling: Foundations and Extensions.* Advanced Quantitative Techniques in the Social Sciences Series, Vol. 10, Sage Newburr Park, CA.

Lee, N., Ridder, G., Strauss, J. (2017). Estimation of poverty transition matrices with noisy data. *Journal of Applied Econometrics*, Vol. 32, No. 1, pp. 37–55.

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

Luttmer, E.F.P. (2002). *Measuring Economic Mobility and Inequality: Disentangling Real Events from Noisy Data*. Harris School of Public Policy, University of Chicago.

Neri, A. (2009). Measuring wealth mobility. *Temi di discussione (Working papers)*, No. 703, Banca d'Italia.

Rulon, P.J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, No. 9, pp. 99-103.

Scott, C. (1992). Estimation of Annual Expenditure from One-month Cross-sectional Data in a Household Survey. *Inter-Stat Bulletin*, No. 8, pp. 57-65.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, Vol. 3, No. 3, pp. 271–295.

Tourangeau, R., Rips, L.J., Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

Victoria-Feser, , M.P. (2000). A General Robust Approach to the Analysis of Income Distribution, Inequality and Poverty. *International Statistical Review*, Vol. 68, pp. 277-293.

Webb, N.M., Shavelson, R.J., Haertel, E.H. (2006). *Reliability Coefficients and Generalizability Theory*. Handbook of Statistics, Vol. 26, Elsevier.