ESTIMATING INCOME POVERTY IN THE PRESENCE OF MEASUREMENT ERROR AND MISSING DATA PROBLEMS

Cheti Nicoletti
(ISER, University of Essex)

Franco Peracchi
Francesca Foliano
(University of Rome "Tor Vergata")

ISER Working Paper
2007-15

Acknowledgement:

We thank Christopher Bollinger, Chuck Manski, Tom Wansbeek and participants to the Workshop "Modeling and Inference in Microeconomics" (Bologna, March 2007) and to the Conference "Measurement Error: Econometrics and Practice" (Birmingham, July 2007) for helpful suggestions. This research was supported by the Economic and Social Research Council through their grant to the Research Centre on Micro-social Change in ISER. Part of this paper is based on work carried out during Francesca Foliano's visit to the European Centre for Analysis in the Social Sciences (ECASS) at the ISER, University of Essex, supported by the Access to Research Infrastructure action under the EU Improving Human Potential Programme.

Readers wishing to cite this document are asked to use the following form of words:

> Nicoletti, Cheti; Peracchi, Franco; Foliano, Francesca (August 2007) 'Estimating Income Poverty in the Presence of Measurement Error and Missing Data Problems', ISER Working Paper 2007-15. Colchester: University of Essex.

The on-line version of this working paper can be found at http://www.iser.essex.ac.uk/pubs/workpaps/

The Institute for Social and Economic Research (ISER) specialises in the production and analysis of longitudinal data. ISER incorporates

- MISOC (the ESRC Research Centre on Micro-social Change), an international centre for research into the lifecourse, and
- ULSC (the ESRC UK Longitudinal Studies Centre), a national resource centre to promote longitudinal surveys and longitudinal research.

The support of both the Economic and Social Research Council (ESRC) and the University of Essex is gratefully acknowledged. The work reported in this paper is part of the scientific programme of the Institute for Social and Economic Research.

Institute for Social and Economic Research, University of Essex, Wivenhoe Park,
Colchester. Essex CO4 3SQ UK
Telephone: +44 (0) 1206 872957 Fax: +44 (0) 1206 873151 E-mail: iser@essex.ac.uk
Website: http://www.iser.essex.ac.uk

ABSTRACT

Reliable measures of poverty are an essential statistical tool to evaluate public policies aimed at reducing poverty. In this paper we consider the reliability of income poverty measures based on survey data which are typically plagued by measurement error and missing data problems. Neglecting these problems can bias the estimated poverty rates. We show how to derive upper and lower bounds for the population poverty rate using only the sample evidence and an upper limit on the probability of misclassifying people into poor and non-poor. By using the European Community Household Panel, we compute bounds for the poverty rate in eleven European countries and study the sensitivity of poverty comparisons across countries to measurement errors and missing data problems.

1 Introduction

Reliable measures of poverty are an essential statistical tool for public policies aimed at reducing poverty. In addition to sampling error problems, estimates of poverty rates from survey data are typically plagued by measurement error and missing data problems. Measurement errors, broadly defined to include editing errors, coding errors, etc., represent the deviations between the recorded answers to a survey question and the underlying attributes being measured. They may reflect systematic misreporting or unreliable response by the interviewee, but may also depend on data collection procedures (questionnaire design and interview methods), the way interviewers interact with the interviewees, and data processing (data entry, editing, coding, etc.). Missing data arise from the failure to obtain a complete response from all individuals included in a survey sample. They may occur because individuals refuse to return their questionnaire (unit non-response) or do not provide an answer for some of the questions (item non-response), and may depend on both individual attitudes and survey procedures. Only relatively recently have statisticians started investigating the impact of these types of non-sampling errors on poverty estimates.

The most common statistical approaches to measurement errors rely on either the classical measurement error model or on mixture models (see van Praag et al. 1983, Ravallion 1994 and Chesher and Schulter 2002 for the classical measurement error model, Cowell and Victoria-Feser 1996 and Pudney and Francavilla 2006 for mixture models, and Bound et al 2001 for a survey of the literature on measurement errors. The former assumes that the observed outcome is equal to the true outcome (the "signal") plus an additive error that has mean zero and is independent of the signal. This strong assumption is often not justified empirically but adopted merely for convenience. A notable case when this assumption is violated is when the outcome is a categorical variable, such as a binary indicator of poverty. On the other hand, mixture models assume that the outcome of interest is mismeasured or a fraction π of individuals, with $0 < \pi < 1$. The observed outcome is then equal to a mixture of two variables, the true outcome and an unknown contaminating variable. The probability of observing the true outcome is $1 - \pi$, while the probability of observing the contaminating variable (the measurement error probability) is π.

For missing data problems, the approaches usually considered by survey methodologists consist of imputation and weighting methods (see Little and Rubin 1987, Rubin 1989, and Rubin 1996). These methods typically assume a missing at random (MAR) condition, that is, they assume

independence between the missing data mechanism and the outcome of interest after conditioning on a set of observed variables. Conversely, econometricians usually adopt methods which take into account selection due to both observed and unobserved variables (see Vella 1998 for a survey). While these methods relax the MAR condition, they usually impose various types of restrictions on the distribution of the unobservables.

Most estimation methods proposed for measurement error or missing data problems focus on point estimation of the parameters of interest, typically at the cost of imposing strong untestable assumptions. Manski and co-authors (see for example Manski 1989 and Horowitz and Manski 1998 for missing data problems, Horowitz and Manski 1995 for measurement error problems, and Manski 2003 for a review of the partial identification approach) have shown how to use the empirical evidence, alone or in conjunction with assumptions that are sufficiently weak to be widely credible, to learn something about the parameters of interest. Their approach involves a shift from point identification to partial identification of the parameters of interest, that is, a shift from the attempt to uncover the "true value" of the parameter of interest to a description of the set of values that are logically possible given the measurement error or missing data mechanisms.

In this paper we follow this partial identification approach and provide bounds on the poverty rate in the presence of both measurement error and missing data problems. We combine results in Nicoletti and Peracchi (2002) and Nicoletti (2003) to bound the poverty rate in the presence of missing data with the approach suggested by Horowitz and Manski (1995) and Molinari (2005) to take measurement errors into account. By using the European Community Household Panel (ECHP), we produce bound estimates for the poverty rates in eleven European countries. The aim of our empirical analysis is to answer the following questions. What can we learn about poverty rates from sample surveys without imposing too restrictive assumptions on the measurement error and the missing data processes? Are bound estimates informative enough to allow us to rank countries in terms of poverty?

The remainder of this paper is organized as follows. Section 2 considers partial identification of poverty rates when the information on poverty status may be affected by measurement error or missing data problems. Section 3 presents the ECHP data that are used in our empirical illustration. Section 4 reports point and bound estimates for the poverty rates using the ECHP data. Finally, Section 5 offers some conclusions.

2 Partial identification of the poverty rate

This section considers partial identification of the poverty rate, conventionally defined as the fraction of people (both children and adults) living in households whose income Y falls below a certain threshold (the "poverty line"). We take Y to be the equivalized net income of a household, that is, its total net income suitably normalized by a measure of the effective size of the household (the "equivalence scale") to allow for possible economies of scale in the use of resources. The poverty line, denoted by γ, is defined in relative terms as a fraction (60 percent) of the median value of equivalized net household income.

Since poverty rates are generally estimated using household survey data, they are typically subject to both sampling and non-sampling errors. Two important sources of non-sampling errors are measurement error and missing data problems (Groves 1989, Biemer et al. 1991, Lessler and Kalsbeek 1992, and Groves et al. 2002). Section 2.1 considers partial identification of the poverty rate in the case of measurement error but no missing data problems. Section 2.2 considers partial identification in the case of missing data but no measurement errors. Finally, Section 2.3 considers partial identification in the case of both measurement errors and missing data.

2.1 Partial identification in the presence of measurement errors

Measurement errors in the poverty status can occur when the household income, the household equivalent scale or the poverty line are measured with errors. In our empirical application, the poverty line is estimated using the imputed values and the weights provided in the ECHP to take into account sampling design and the presence of fully non-responding households, unit non-response within responding households and item non-response. Hence, it may be affected by both sampling noise and systematic bias.

If we denote by D_Y the true (error-free) but possibly unobserved poverty indicator, which is equal to one if $Y \leq \gamma$ and is equal to zero otherwise, then the true poverty rate is just the probability of being poor, $\Pr(D_Y = 1) = \Pr(Y \leq \gamma)$. If we denote by W the observed (error-ridden) equivalized net household income and by $\widehat{\gamma}$ the estimated poverty line, then the observed poverty indicator D_W is equal to one if $W \leq \widehat{\gamma}$ and is equal to zero otherwise, and the observed poverty rate is $\Pr(D_W = 1) = \Pr(W \leq \widehat{\gamma})$. When $D_Y \neq D_W$, poverty status is measured with error. Since both D_Y and D_W are categorical indicators, the measurement error problem is a problem of

misclassification. The problem may arise either because $Y \neq W$ or because $\widehat{\gamma} \neq \gamma$ due to sampling noise or systematic bias. Ignoring the problem may lead to biased estimates of the true poverty rate $\Pr(D_Y = 1)$. As an alternative we consider using the sample information to partially identify the true poverty rate. This involves finding non-trivial upper and lower bounds for the set of values that $\Pr(D_Y = 1)$ can take given the measurement error process. This is the approach adopted by Horowitz and Manski (1995), Molinari (2005) and Pudney and Francavilla (2006). In the following, we give some details on their methods and then describe our proposal.

Horowitz and Manski (1995) show how to identify bounds for parameters of interest by considering a mixture model and assuming the existence of a non-trivial upper bound on the probability of error. Although their results are quite general, analytical expressions for the bounds are only available in a few special cases. In general, bounds have to be computed by solving a non-linear constrained optimization problem.

Chavez-Martin del Campo (2004) specializes the results of Horowitz and Manski (1995) to the case of poverty measures. By considering a mixture model for household income and assuming a non-trivial upper bound on the measurement error probability, he shows how to bound poverty measures that are additively separable. The poverty measure used in this paper, namely the fraction of people below 60% of the equivalized median income, belongs to this class.

An alternative approach is to bound the poverty rate directly by considering a mixture model for the poverty indicator rather than for income. The main advantage of this approach is that we can take into account all the errors which may lead to misclassifying poverty status–errors affecting the income measure, the equivalence scale, or the poverty line–without having to explicitly model their role. Since the poverty indicator is binary, the mixture model takes the form

$$D_W = D_Y(1 - Z) + D_V Z, \tag{1}$$

where D_V is an erroneous poverty indicator and Z is a binary indicator equal to one if there is misclassification (that is, $D_W \neq D_Y$) and equal to zero otherwise. Because $D_W = D_Y$ if $Z = 0$ and $D_W = D_V$ if $Z = 1$, the observed poverty rate may be written

$$\Pr(D_W = 1) = \Pr(D_Y = 1 \,|\, Z = 0)\ \Pr(Z = 0) + \Pr(D_V = 1 \,|\, Z = 1)\ \Pr(Z = 1), \tag{2}$$

where $\Pr(Z = 1) = \Pr(D_Y \neq D_W)$ is the probability of misclassification, or measurement error probability. The mixture model (1) is an example of contaminated sampling model if measurement

error and poverty status are independent, i.e. $\Pr(D_Y \mid Z = 0) = \Pr(D_Y)$, and is an example of corrupted sampling model if the independence condition does not hold.

It turns out that assuming a corrupted sampling model imposes no restriction on the relationship between an error-ridden indicator and the underlying error-free indicator. To see this, notice that if there is misclassification ($Z = 1$), then $D_V = 1$ if and only if $D_Y = 0$. Therefore, we can rewrite equation (2) as

$$\Pr(D_W = 1) = \Pr(D_Y = 1 \mid Z = 0)\ \Pr(Z = 0) + \Pr(D_Y = 0 \mid Z = 1)\ \Pr(Z = 1).$$

Since

$$\Pr(D_Y = 1 \mid Z = 0)\ \Pr(Z = 0) = \Pr(D_Y = 1, D_W = 1)$$

and

$$\Pr(D_Y = 0 \mid Z = 1)\ \Pr(Z = 1) = \Pr(D_Y = 0, D_W = 1),$$

we can further rewrite (2) as

$$\Pr(D_W = 1) = \Pr(D_W = 1 \mid D_Y = 1)\ \Pr(D_Y = 1) + \Pr(D_W = 1 \mid D_Y = 0)\ \Pr(D_Y = 0). \qquad (3)$$

By a similar argument, the observed probability of being non-poor can be written as

$$\Pr(D_W = 0) = \Pr(D_W = 0 \mid D_Y = 0)\ \Pr(D_Y = 0) + \Pr(D_W = 0 \mid D_Y = 1)\ \Pr(D_Y = 1). \qquad (4)$$

Because (3) and (4) are implied by the law of total probability, the corrupted sampling model imposes no restrictions on the relationship between an error-ridden and an error-free variable when both are binary indicators. This result may be generalized to the case of a categorical variable.

Equations (3) and (4) correspond to what Molinari (2005) calls the "direct misclassification model", adapted here to the case of a binary indicator. Molinari's paper shows how to identify bounds for the distribution of an unobserved categorical variable by imposing various assumptions on the direct misclassification model (3)–(4). In this paper, we focus on two of these assumptions:

Assumption A1 $\sum_{j=0}^{1} \Pr(D_W = j, D_Y = j) \geq 1 - \lambda > 0$.

Assumption A2 $\Pr(D_W = j \mid D_Y = j) \geq 1 - \lambda > 0$ for $j = 0, 1$.

Because $\sum_{j=0}^{1} \Pr(D_W = j, D_Y = j) = \Pr(D_W = D_Y)$, Assumption A1 is equivalent to the assumption that $\Pr(D_W \neq D_Y) < \lambda < 1$. This assumption restricts the joint distribution of D_Y and

D_W by imposing a non-trivial upper bound on the probability of misclassification. Assumption A2 implies Assumption A1 and is therefore stronger. In some applications, it may be possible to directly estimate the upper bound λ in Assumptions A1 and A2. As pointed out by Horowitz and Manski (1995), even when this is not possible, it may still be of interest to determine how inference on population parameters changes with changes in λ.

Proposition 3 in Molinari (2005) shows that Assumption A1 implies the following upper and lower bounds on the true poverty rate

$$\text{UB}_{A1} = \min\{\Pr(D_W = 1) + \lambda,\, 1\},$$
$$\text{LB}_{A1} = \max\{\Pr(D_W = 1) - \lambda,\, 0\}.$$

If $\lambda \leq \Pr(D_W = 1) \leq 1-\lambda$, then the width of the resulting identification region for the true poverty rate is equal to 2λ. Assumption A2 implies instead the following bounds

$$\text{UB}_{A2} = \min\left\{\frac{\Pr(D_W = 1)}{1-\lambda},\, 1\right\},$$
$$\text{LB}_{A2} = \max\left\{\frac{\Pr(D_W = 1) - \lambda}{1-\lambda},\, 0\right\}.$$

If $\lambda \leq \Pr(D_W = 1) \leq 1-\lambda$, then the width of the resulting identification region for the true poverty rate is equal to $\lambda/(1-\lambda)$. Notice that, although stronger than Assumption A1, Assumption A2 implies tighter bounds only when $\lambda < .5$, at least if we again assume that $\lambda \leq \Pr(D_W = 1) \leq 1-\lambda$.

Pudney and Francavilla (2006) investigate the effect of measurement errors on the estimation of poverty rates by considering a contaminated sampling model for household income rather than for the poverty indicator, and by conditioning the probability of being poor on an indicator of deprivation. Thanks to assumptions on the relationship between poverty status and the indicator of deprivation, and to additional independence assumptions, they prove that it is possible to exactly identify the poverty rate. They also show how to obtain partial identification of the poverty rate when some of these assumptions are relaxed.

Although our approach is similar in spirit to those just outlined, our starting point is neither the mixture model (1) nor the direct misclassification model (3)–(4). Instead, we consider the following indirect misclassification model

$$\begin{bmatrix} \Pr(D_Y = 1) \\ \Pr(D_Y = 0) \end{bmatrix} = \begin{bmatrix} \Pr(D_Y = 1 \,|\, D_W = 1) & \Pr(D_Y = 1 \,|\, D_W = 0) \\ \Pr(D_Y = 0 \,|\, D_W = 1) & \Pr(D_Y = 0 \,|\, D_W = 0) \end{bmatrix} \begin{bmatrix} \Pr(D_W = 1) \\ \Pr(D_W = 0) \end{bmatrix}. \tag{5}$$

This model maps $\Pr(D_W = 1)$ and $\Pr(D_W = 0)$, which are point-identified by the sampling process, into $\Pr(D_Y = 1)$ and $\Pr(D_Y = 0)$, which are not point-identified. Notice that, just like the direct

misclassification model (3)–(4), the indirect misclassification model (5) is simply an implication of the law of total probability and imposes no restriction on the relationship between the error-free and the error-ridden indicator of poverty.

To partially identify the true poverty rate, we consider the following two assumptions:

Assumption B1 $\sum_{j=0}^{1} \Pr(D_W = j, D_Y = j) \geq 1 - \lambda > 0$.

Assumption B2 $\Pr(D_Y = j \,|\, D_W = j) \geq 1 - \lambda > 0$ for $j = 0, 1$.

Assumption B1 is the same as Assumption A1. Unlike Molinari's Assumption A2, which implies an upper bound on the direct misclassification probabilities $\Pr(D_W = j \,|\, D_Y = i)$, for $i \neq j$, Assumption B2 implies an upper bound on the indirect misclassification probabilities $\Pr(D_Y = j \,|\, D_W = i)$, for $i \neq j$. Thus, while Assumption A2 restricts the conditional distribution of D_W given D_Y, Assumption B2 restricts the conditional distribution of D_Y given D_W.

The next proposition presents the bounds on the true poverty rate implied by Assumptions B1 and B2. All proofs are collected in Appendix B).

Proposition 1 *If Assumption B1 holds, then*

$$\mathrm{UB}_{B1} = \mathrm{UB}_{A1} = \min\{\Pr(D_W = 1) + \lambda,\, 1\},$$
$$\mathrm{LB}_{B1} = \mathrm{LB}_{A1} = \max\{\Pr(D_W = 1) - \lambda,\, 0\}.$$

If Assumption B2 holds, then

$$\mathrm{UB}_{B2} = \Pr(D_W = 1) + \lambda \Pr(D_W = 0),$$
$$\mathrm{LB}_{B2} = \Pr(D_W = 1)(1 - \lambda).$$

Further, all these bounds are sharp.

Under Assumption B1, the indirect classification model gives the same bounds as the direct classification model under Assumption A1. Under Assumption B2, the width of the identification region for the true poverty rate is equal to λ. Because Assumption B2 implies Assumption B1, the fact that it gives a narrower identification region is not surprising. Perhaps more surprising is the fact that the indirect misclassification model under Assumption B2 implies tighter bounds than the direct misclassification model under Assumption A2.

2.2 Partial identification in the presence of missing data

We now consider the case when there are no measurement errors but, because of unit or item nonresponse, income data are missing for a fraction of the households. Following Manski (1989), let D_R be a binary indicator equal to one if an individual belongs to a responding household, namely one whose income is fully reported and equal to zero otherwise. If γ is the poverty line then, by the law of total probability, the true poverty rate satisfies

$$\Pr(Y \leq \gamma) = \Pr(Y \leq \gamma \,|\, D_R = 1)\ \Pr(D_R = 1) + \Pr(Y \leq \gamma \,|\, D_R = 0)\ \Pr(D_R = 0). \tag{6}$$

Because only three of the four elements on the right hand side of (6) can be identified from the sampling process, the true poverty rate is not point-identified unless additional assumptions are made. However, because the unknown element $\Pr(Y \leq \gamma \,|\, D_R = 0)$ is bounded between zero and one, substituting its maximum and minimum values in (6) gives the following upper and lower bounds on the true poverty rate

$$\begin{aligned} \text{UB} &= \Pr(Y \leq \gamma \,|\, D_R = 1)\ \Pr(D_R = 1) + \Pr(D_R = 0), \\ \text{LB} &= \Pr(Y \leq \gamma \,|\, D_R = 1)\ \Pr(D_R = 1). \end{aligned}$$

The width of the resulting identification region is equal to the non-response probability $\Pr(D_R = 0)$. It is straightforward to prove that these bounds are sharp.

An important question is how to narrow these "worst-case" bounds, that is, how to sharpen our inference by reducing the range of plausible values for the poverty rate. We begin by noticing that many non-respondents provide partial information on their income. For example, household income is typically obtained by adding a number of income components (wages and salaries, self-employment income, pensions, etc.) across all household members. Usually, nonresponse at the household level is only partial, in the sense that at least some household members provide information on at least some of the income components that they received.

If Y^* denotes partially reported income, that is, the sum of all reported income components across all members of the household, then the unknown poverty rate among the non-respondents may be decomposed as follows

$$\begin{aligned} \Pr(Y \leq \gamma \,|\, D_R = 0) = \Pr(Y \leq \gamma \,|\, Y^* \leq \gamma, D_R = 0)\ \Pr(Y^* \leq \gamma \,|\, D_R = 0) \\ + \Pr(Y \leq \gamma \,|\, Y^* > \gamma, D_R = 0)\ \Pr(Y^* > \gamma \,|\, D_R = 0), \end{aligned} \tag{7}$$

where, in the absence of measurement errors, $\Pr(Y \leq \gamma \,|\, Y^* > \gamma, D_R = 0) = 0$ because partially reported income Y^* cannot exceed true income Y. Since the probability $\Pr(Y \leq \gamma \,|\, Y^* \leq \gamma, D_R = 0)$ must necessary lie between zero and one, we obtain the following upper and lower bounds

$$\text{UB}^* = \Pr(Y \leq \gamma \,|\, D_R = 1)\ \Pr(D_R = 1) + \Pr(Y^* \leq \gamma \,|\, D_R = 0)\ \Pr(D_R = 0),$$
$$\text{LB}^* = \text{LB} = \Pr(Y \leq \gamma \,|\, D_R = 1)\ \Pr(D_R = 1).$$

Thus, the information on partially reported income provides a sharper upper bound on the poverty rate but does not affect the lower bound, which remains the same as the worst case bound LB. This narrows the width of the identification region from $\Pr(D_R = 0)$ to $\Pr(Y^* \leq \gamma \,|\, D_R = 0)\ \Pr(D_R = 0)$. Our use of partially reported income to narrow the Manski bounds is new, but is to some extent a modification of the methods proposed by Vasquez-Alvarez et al. (1999, 2001) and Manski and Tamer (2002) for interval data.

2.3 Partial identification in the presence of measurement errors and missing data

In the presence of both measurement errors and missing data, identification of the poverty rate becomes more difficult. In the equation

$$\Pr(D_Y = 1) = \Pr(D_Y = 1 \,|\, D_R = 1)\ \Pr(D_R = 1) + \Pr(D_Y = 1 \,|\, D_R = 0)\ \Pr(D_R = 0),$$

both $\Pr(D_Y = 1 \,|\, D_R = 1)$ and $\Pr(D_Y = 1 \,|\, D_R = 0)$ are now unknown. This is because for responding people we observe an erroneous poverty indicator D_W instead of the true indicator D_Y, while for non-responding people we observe neither D_W nor D_Y.

Although misclassification of poverty status may affect both respondents and non-respondents, in practice the problem is relevant only for people whose household income is observed. Therefore, the partial identification methods discussed in Section 2.1 can be directly applied to find a lower and an upper bound for $\Pr(D_Y = 1 \,|\, D_R = 1)$, the poverty rate for respondents. All we need is an upper bound on either the measurement error probability, the direct misclassification probabilities, or the indirect misclassification probabilities, after conditioning on the event $D_R = 1$. Thus consider the following assumptions:

Assumption C1 $\sum_{j=0}^{1} \Pr(D_W = j, D_Y = j \,|\, D_R = 1) \geq 1 - \lambda > 0$;

Assumption C2 $\Pr(D_Y = j \,|\, D_W = j, D_R = 1) \geq 1 - \lambda > 0$ for $j = 0, 1$;

Assumption C3 $\Pr(D_W = j \,|\, D_Y = j, D_R = 1) \geq 1 - \lambda > 0$ for $j = 0, 1$.

The next proposition gives, for each of these three assumptions, the implied bounds on the true poverty rate.

Proposition 2 *If Assumption C1 holds, then*

$$\begin{aligned} \mathrm{UB}_{C1} &= \min\{\Pr(D_W = 1 \,|\, D_R = 1) + \lambda, 1\} \Pr(D_R = 1) + \Pr(D_R = 0), \\ \mathrm{LB}_{C1} &= \max\{\Pr(D_W = 1 \,|\, D_R = 1) - \lambda, 0\} \Pr(D_R = 1). \end{aligned} \tag{8}$$

If Assumption C2 holds, then

$$\begin{aligned} \mathrm{UB}_{C2} &= [\Pr(D_W = 1 \,|\, D_R = 1) + \lambda \Pr\{D_W = 0 \,|\, D_R = 1\}] \Pr(D_R = 1) + \Pr(D_R = 0), \\ \mathrm{LB}_{C2} &= \Pr(D_W = 1 \,|\, D_R = 1)(1 - \lambda) \Pr(D_R = 1). \end{aligned} \tag{9}$$

If Assumption C3 holds, then

$$\begin{aligned} \mathrm{UB}_{C3} &= \min\{\Pr(D_W = 1 \,|\, D_R = 1)/(1 - \lambda), 1\} \Pr(D_R = 1) + \Pr(D_R = 0), \\ \mathrm{LB}_{C3} &= \max\{[\Pr(D_W = 1 \,|\, D_R = 1) - \lambda]/(1 - \lambda), 0\} \Pr(D_R = 1). \end{aligned} \tag{10}$$

Further, all these bounds are sharp.

Molinari (2005) suggests taking into account both missing data and measurement error problems by considering an extended direct misclassification model. However, she does not derive an analytical expression for the bounds. In the case of a binary indicator, she suggests considering an error-ridden variable which takes on three possible values: zero, one or missing. We prefer to consider three events which are mutually exclusive and collectively exhaustive: $D_W = 1$ and $D_R = 1$ (reporting poor), $D_W = 0$ and $D_R = 1$ (reporting non-poor), and $D_R = 0$ (partially responding household). We then extend the indirect misclassification model (5) by defining the true poverty probability as a linear function of the above three events. The main advantage of our approach is that derivation of the bounds is much easier. In particular, we are able to obtain an upper and a lower bound for the true poverty rate by simply assuming an upper and a lower bound for the poverty probability of respondents and non-respondents. In Appendix A we give details on this extended indirect misclassification model and possible extensions to interval data.

If non-respondents provide partial information, then we can narrow the bounds further by using the information on partially reported income Y^* as shown in Section 2.2. In this case, the term $\Pr(D_R = 0)$ in each of the upper bounds UB_{C1}, UB_{C2} and UB_{C3} may be replaced by

$\Pr(D_R = 0)\ \Pr(Y^* \leq \gamma \,|\, D_R = 0)$. Estimating these new bounds requires a knowledge of the true partially reported income Y^* and the true poverty line γ. If we only observe W^* and $\widehat{\gamma}$, which are error-ridden measurements of Y^* and γ, then we must modify equation (7) as follows

$$\begin{aligned}\Pr(Y \leq \gamma \,|\, D_R = 0) = {} & \Pr(Y \leq \gamma \,|\, D_{W^*} = 1, D_R = 0)\ \Pr(D_{W^*} = 1 \,|\, D_R = 0) \\ & + \Pr(Y \leq \gamma \,|\, D_{W^*} = 0, D_R = 0)\ \Pr(D_{W^*} = 0 \,|\, D_R = 0),\end{aligned} \tag{11}$$

where D_{W^*} is equal to one if $W^* \leq \widehat{\gamma}$ and is equal to zero otherwise. In the absence of measurement errors we could safely assume that

$$\Pr(Y \leq \gamma \,|\, D_{W^*} = 0, D_R = 0) = \Pr(Y \leq \gamma \,|\, Y^* > \gamma, D_R = 0) = 0.$$

In the presence of measurement errors, however, $\Pr(Y \leq \gamma \,|\, D_{W^*} = 1, D_R = 0)$ can be greater than zero. For this reason we introduce the following additional assumption:

Assumption C4 $\Pr(D_Y = 1 \,|\, D_{W^*} = 0, D_R = 0) \leq \delta < \lambda$.

Assuming that $\delta < \lambda$ is plausible because, if the true income is below the poverty line, then it seems unlikely for the observed partially reported income to exceed the observed poverty line.

Combining Assumptions C1 and C4 into Assumption C1*, it is easy to modify the argument in Proposition 2 to derive the following upper and lower bounds on the poverty rate

$$\begin{aligned}\mathrm{UB}_{C1^*} = {} & \min\{\Pr(D_W = 1 \,|\, D_R = 1) + \lambda, 1\}\ \Pr(D_R = 1) \\ & + \Pr(D_R = 0)\,[\Pr(D_{W^*} = 1 \,|\, D_R = 0)(1 - \delta) + \delta],\end{aligned} \tag{12}$$

$$\mathrm{LB}_{C1^*} = \mathrm{LB}_{C1} = \max\{\Pr(D_W = 1 \,|\, D_R = 1) - \lambda, 0\}\ \Pr(D_R = 1).$$

Similarly, combining Assumptions C2 and C4 into Assumption C2* gives

$$\begin{aligned}\mathrm{UB}_{C2^*} = {} & [\Pr(D_W = 1 \,|\, D_R = 1) + \lambda \Pr\{D_W = 0 \,|\, D_R = 1\}]\ \Pr(D_R = 1) \\ & + \Pr(D_R = 0)\,[\Pr(D_{W^*} = 1 \,|\, D_R = 0)(1 - \delta) + \delta],\end{aligned} \tag{13}$$

$$\mathrm{LB}_{C2^*} = \mathrm{LB}_{C2} = \Pr(D_W = 1 \,|\, D_R = 1)(1 - \lambda)\ \Pr(D_R = 1).$$

Finally, combining Assumptions C3 and C4 into Assumption C3* gives

$$\begin{aligned}\mathrm{UB}_{C3^*} = {} & \min\{\Pr(D_W = 1 \,|\, D_R = 1)/(1 - \lambda), 1\}\ \Pr(D_R = 1) \\ & + \Pr(D_R = 0)\,[\Pr(D_{W^*} = 1 \,|\, D_R = 0)(1 - \delta) + \delta],\end{aligned} \tag{14}$$

$$\mathrm{LB}_{C3^*} = \mathrm{LB}_{C3} = \max\{[\Pr(D_W = 1 \,|\, D_R = 1) - \lambda]/(1 - \lambda), 0\}\ \Pr(D_R = 1).$$

Assumption C4 and information on the observed reported income cause the various identification regions to shrink by an amount equal to $\Pr(D_R = 0)\,[1 - \delta - (1 - \delta)\Pr(D_{W^*} = 1 \,|\, D_R = 0)]$. Again, it is easy to show that all the above bounds are sharp.

3 Data

We now describe the data set used in our empirical illustration, namely the European Community Household Panel (ECHP), and the construction of our poverty indicator.

The ECHP is a longitudinal survey of households and individuals, centrally designed and coordinated by the Statistical Office of the European Communities (Eurostat) and conducted annually from 1994 to 2001. Its target population consists of all individuals living in private households within the European Union. In its first wave (1994), the survey covered about 60,000 households and 130,000 individuals in twelve countries, namely Belgium, Denmark, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain and the UK. Austria, Finland and Sweden began to participate to the ECHP only later, respectively from the second (1995), third (1996) and fourth (1997) wave. In Belgium and the Netherlands, the ECHP was linked from the beginning to already existing national panels. In Germany, Luxembourg and the UK, instead, the first three waves of the ECHP ran parallel to already existing national panels, respectively the German Social Economic Panel (GSOEP), the Luxembourg Social Economic Panel (PSELL) and the British Household Panel Survey (BHPS). Starting from the fourth (1997) wave, it was decided to merge the ECHP into the GSOEP, the PSELL and the BHPS.

In this paper, we focus attention on the eleven countries who participated in the survey for the whole period 1994–2001, namely Belgium, Denmark, France, Germany, Greece, Ireland, Italy, Netherlands, Portugal, Spain, and the UK. Our sample consists of 116,354 individuals observed in the most recent wave of the panel (2001). We include in the sample only households for which at least the reference person returned the household questionnaire. The percentage of fully nonresponding households (households for which neither the household questionnaire nor any of the personal questionnaires was returned) is quite small, on average 2 percent in each wave. We take into account this type of nonresponse and sampling design by using the weights provided in the public-use files of the ECHP.

The key variable for the construction of our poverty indicator is total net household income, which is the sum of all incomes (wages and salaries, self-employment income, pensions, etc.) received by all members of a given household. Nonresponse to this variable may occur either because some household members do not return their personal questionnaire or because household members returning their personal questionnaire do not answer all income questions. Excluding from the

sample fully nonresponding housheolds, problems of complete or partial nonresponse to household income arise for 15,454 individuals (about 13.3 percent of our sample). Table 1 shows nonresponse rates by country. Belgium, Italy and Germany present the highest nonresponse rates (above 15 percent), while France, Ireland and the Netherlands are the countries where nonresponse rates are smallest (between 7 and 10 percent).

Household income provided in the ECHP is the annual amount in the year before the survey, net of taxes and expressed in national currency units and current prices. We then divide real household income by the modified OECD equivalence scale to take household size and composition into account.

4 Empirical results

This section presents our empirical results using the ECHP data. Section 4.1 presents point estimates that use imputed income to take missing data problems into account but neglect measurement error problems. Section 4.2 presents the estimated bounds on true poverty rates taking both missing data and measurement error problems into account.

4.1 Point estimates

Table 1 shows, for each of the eleven European countries considered, unweighted and weighted point estimates of the poverty rates for the overall population and their estimated standard errors (in parenthesis). As with most official statistics on poverty, we take missing data problems into account by using imputed values but we neglect measurement error problems.

Unweighed poverty rates are defined as the fraction of the population with (imputed or observed) equivalized household income below 60% of the national median of equivalized household income. Weighted poverty rates are computed in the same way, but using the survey weights provided by the ECHP to take sampling design and full household non-response into account.

Using weighted or unweighted estimates somewhat changes the ranking of countries. Comparing the two rankings, we can identify three groups of countries: the first group consists of the countries with the lowest poverty rates (always below 15 percent), namely Belgium, Denmark, Germany and the Netherlands; the second group consists of the middle-ranking countries, namely France, Ireland, Spain and the UK (always between 15 and 20 percent); the third group consists of Greece, Italy, and Portugal with the highest poverty rates (always above 20 percent). This grouping largely agrees

with those reported by Eurostat publications, based either on the ECHP or, for the most recent years, on the Community Statistics on Income and Living Conditions or EU-SILC (see for example Dennis and Guio 2004 and Guio 2004). The main difference concerns Ireland, which appears at the bottom of the ranking of poverty rates calculated using the EU-SILC.

4.2 Bound estimates

This section presents estimated bounds for the true poverty rates based on the theoretical bounds in Section 2.3. These bounds are functions of probabilities which are non-parametrically estimated by simple weighted empirical frequencies, using the weights provided in the ECHP. Since the bounds are estimated, we also take their sampling variability into account. This is done by constructing, for each bound, 95%-level bootstrap confidence intervals based on the percentile method and 1,000 bootstrap replications. Unlike standard asymptotic confidence intervals, these confidence intervals are in general not symmetric. The bootstrap samples are obtained by sampling household (not individuals) with replacement. Further, for each bootstrap sample, the cross-sectional weights are rescaled to keep their mean equal to one (for more details on bootstrap inference for poverty measures see Biewen 2002).

Table 2 reports, separately by country, the estimated upper and lower bounds for the poverty rates and the corresponding upper and lower limits of their bootstrap confidence interval, respectively in the first and second row for each country. The results are reported separately for the three assumptions C1*, C2* and C3*, which identify the following intervals $[\text{LB}_{C1^*}, \text{UB}_{C1^*}]$, $[\text{LB}_{C2^*}, \text{UB}_{C2^*}]$, and $[\text{LB}_{C3^*}, \text{UB}_{C3^*}]$. We take $\lambda = .10$ as the upper bound for the measurement error probability and for the direct and indirect misclassification probabilities. The upper bound ρ for the probability that a person is poor when his/her partially reported household income is higher than the estimated poverty line is set equal to .025 (one fourth of λ). These choices are based on the validation studies quoted by Horowitz and Manski (1995) and Molinari (2005), who suggest that error probabilities usually range between .01 and .10, and on the study of Epland and Kirkeberg (2002), who compare true and reported Norwegian income data by matching administrative registers with the Survey of Living Conditions in 1996. Using Table 2 in Epland and Kirkeberg (2002), and setting the poverty line at 50,000 Norwgian crowns, we obtain the following results: (i) the probability that the true and the observed poverty status are different is 3 percent; (ii) the probability that the true poverty status is one (zero) given that the observed status is zero (one)

is 2 percent (6 percent); (iii) the probability that the observed poverty status is one (zero) given that the true status is zero (one) is 1 percent (11 percent). From these results, it is reasonable to put $\lambda = .10$.

We also carried out a sensitivity analysis using different values of λ and ρ. Table 2 shows that the intervals $[\text{LB}_{C2^*}, \text{UB}_{C2^*}]$ and $[\text{LB}_{C3^*}, \text{UB}_{C3^*}]$ partially overlap, and are both contained in the larger interval $[\text{LB}_{C1^*}, \text{UB}_{C1^*}]$. Assumption C2* generally produces higher lower bounds than Assumption C3*, while the opposite is true for upper bounds.

If we assume that Assumptions C2* and C3* both hold, then we can compute narrower bounds, which we call $[\text{LB}_D, \text{UB}_D]$. The upper bound UB_D is given by the minimum between UB_{C2^*} and UB_{C3^*}, while the lower bound LB_D is given by the maximum between LB_{C2^*} and LB_{C3^*}. Estimates of this new set of bounds are presented in Table 3.

The relationship between the width of the identification regions under different sets of assumptions are in line with the theoretical results. The interval $[\text{LB}_{C1^*}, \text{UB}_{C1^*}]$ is the widest, followed (in the order) by the intervals $[\text{LB}_{C2^*}, \text{UB}_{C2^*}]$, $[\text{LB}_{C3^*}, \text{UB}_{C3^*}]$ and $[\text{LB}_D, \text{UB}_D]$. The width of these intervals measures how serious the identification problem is. A zero width corresponds to point identification of the poverty rate. A width that is positive but less than one corresponds to partial identification of the poverty rate. Finally, a width that is equal to one implies complete lack of identification.

Table 3 reports the width of the narrowest interval $[\text{LB}_D, \text{UB}_D]$. More precisely, the table reports, separately by country, the width measured by considering the estimated lower and upper bounds (first row) and the corresponding estimated lower and upper bootstrap confidence limits (second row). The width varies between .055 (.086 in terms of bootstrap confidence limits) and .118 (.209 in terms of bootstrap confidence limits). This implies that, although point identification is impossible, the range of plausible values for the poverty rate can be bounded by a narrow interval whose width is equal to .118 (or .209) in the worst case.

Table 3 also shows a decomposition of the width of the interval $[\text{LB}_D, \text{UB}_D]$ into two additive components, denoted by W_1 and W_2. The first component,

$$W_1 = \Pr(D_R = 0)\,[\Pr(D_{W^*} = 1 \,|\, D_R = 0)(1 - \delta) + \delta],$$

is caused by the presence of missing data, and the fact that $\Pr(D_R = 0) > 0$ and $\Pr(D_Y =$

$1 \mid D_{W^*} = 0, D_R = 0) > 0$. The second component,

$$W_2 = \mathrm{UB}_D - \mathrm{LB}_D - W_1 = W - W_1,$$

is instead caused by measurement errors affecting the observed poverty indicator. A similar decomposition can be used when considering the sets of assumptions C1*, C2* and C3*; the first component is the same for each assumption, while the second component depends on the particular assumption considered. In all countries except France, at most 40 percent of the interval width (53.4 percent in terms of bootstrap confidence limits) is due to measurement error problems. This suggests that lack of identification is mainly caused by missing data problems, at least when assuming $\lambda = 0.10$ and $\rho = 0.025$.

We also conducted a sensitivity analysis by assuming $\rho = .25\lambda$ and considering different values of λ ranging between .010 and .990. Table 4 reports the minimum and the maximum widths across country of the estimated interval $[\mathrm{LB}_D, \mathrm{UB}_D]$. Both the minimum and the maximum widths increase with λ. The widths are always smaller than .25 for values of λ less or equal to .30. The widths become instead quite large when λ exceeds .50. The interval width is mainly explained by measurement error problems if $\lambda \geq 0.30$, and by missing data problems if $\lambda \leq .10$. These results may be useful to survey methodologists interested in improving the quality of a survey by adopting techniques aimed at reducing either non-response rates or measurement errors.

In another sensitivity analysis, we keep λ fixed at .10 and consider varying ρ in the range from .025 to .99. The interval width increases only slightly with increasing ρ. Even for a value of ρ as high as .99, the minimum width is .099 and the maximum width is .230.

The estimated identification regions for the poverty rate overlap partially across countries, except when we compare Denmark with Greece, Italy, Portugal, Ireland and the UK. By ranking countries in terms of their lower or upper bound under Assumptions C2, C3 or C4, we are able to identify three groups of countries: Belgium, Denmark, Germany and the Netherlands belong to the low-poverty group, Greece, Italy and Portugal belong to the high-poverty group, while France, Ireland, Spain and the UK make up an intermediate group. Interestingly, this is exactly the country ranking obtained using the point estimates of poverty rates in Section 4.1. Our bound estimates seem also in line with the official statistics published by Eurostat, where Greece, Italy and Portugal are the countries with the highest poverty rates, around 23 percent, and Belgium, Denmark, Germany and the Netherlands have the lowest poverty rates, usually lower than 15 percent. In

fact, the estimated lower bounds for the poverty rate are always higher than 15 percent in Greece, Italy and Portugal, while the estimated upper bounds are always lower than 23 percent in Belgium, Denmark, Germany and the Netherlands.

5 Conclusions

In this paper we suggest new ways of partially identifying poverty rates in the presence of both measurement error and missing data problems. We show that one can analytically compute bounds for the poverty rates by assuming the existence of a non-trivial upper bound on one of the following three probabilities: (i) the probability that the observed and the true poverty indicators are different, or measurement error probability, (ii) the probability that the true poverty indicator is one (zero) given that the observed poverty indicator is zero (one), or direct misclassification probability, or (iii) the probability that the observed poverty indicator is one (zero) given that the true indicator is zero (one), or indirect misclassification probability.

An upper bound on the probability of measurement errors is the main assumption imposed by Horowitz and Manski (1995) who consider the corrupted sampling model, while upper bounds on either the probability of measurement errors or the direct misclassification probabilities are the main assumptions imposed by Molinari (2005), who considers the direct misclassification model. Notice that, in our case, the corrupted sampling model is equivalent to the direct misclassification model. Furthermore, both models are implications of the law of total probability and therefore impose no restriction on the relationship between the observed (error-ridden) and the true (error-free) outcome. In this paper we introduce the indirect misclassification model, which is also an implication of the law of total probability and allows us to derive bounds on the poverty rate under the assumption of an upper bound on the indirect misclassification probabilities. We show that the indirect misclassification approach can be extended to the case where measurement error and missing data (or interval data) problems coexist, and we compute analytical expressions for the bounds which can be easily estimated non-parametrically.

Using the indirect misclassification approach, we show that it is possible to say something meaningful about the ranking of European countries in terms of poverty rates. An unambiguous ranking is not possible, however, because estimated intervals for poverty rates partially overlap across countries, except when comparing Denmark with Greece, Italy, Portugal, Ireland and the

UK. The lack of identification is due more to the presence of missing data than to measurement errors, at least when we set the upper bound to .10 for the probability of measurement error and the indirect and direct misclassification probabilities.

Interestingly, our bound estimates are in line with the official statistics published by Eurostat, where Greece, Italy and Portugal are the countries with the highest poverty rates, around 23 percent, while Belgium, Denmark, Germany and the Netherlands have the lowest poverty rates, usually lower than 15 percent. In fact, our estimated lower bounds for the poverty rate are higher than 15 percent in Greece, Italy and Portugal, while the estimated upper bounds are lower than 23 percent in Belgium, Denmark, Germany and the Netherlands.

Possible extensions for future research involve the use of additional assumptions to further tighten the bounds. For example, one may consider instrumental variables and monotone instrumental variables assumptions (as in Manski and Pepper, 2000), verification restrictions (as in Dominitz and Sherman 2006, and Kreider and Pepper 2007), or restrictions suggested by economic theory (as in Blundell et al 2007). However, in a cross-country comparison of poverty it is not easy to find assumptions which are credible and equally plausible for all countries.

References

Biemer, P.P., Groves R.M., Lyberg L.E., Mathiowetz N.A., Sudman S. (1991), *Measurement Errors in Surveys*, Wiley, New York.

Biewen M. (2002), Bootstrap inference for inequality, mobility and poverty measurement, *Journal of Econometrics*, 108: 317–342.

Blundell R., Gosling A., Ichimura H., Meghir C. (2007), Changes in the distribution of male and female wages accounting for employment composition using bounds, *Econometrica*, 75: 323–363.

Bound J, Brown C., Mathiowetz N. (2001), Measurement error in survey data, in J.J. Heckamn and E. Leamer (eds.), *Handbook of Econometrics,* , vol. 5, North Holland, 3705–3843.

Chavez-Martin del Campo J.C. (2004), Partial identification of poverty measures with contaminated data, mimeo, Econometric Society 2004 Latin American Meetings.

Chesher A., Schulter C. (2002), Welfare measurement and measurement error, *Review of Economic Studies*, 69: 357–378.

Cowell F.A., Victoria-Feser M.-P. (1996), Robustness properties of inequality measures, *Econometrica*, 64: 77–101.

Dennis I., Guio A.C. (2004), Poverty and social exclusion in the EU, *Population and Social Conditions*, 16/2004, Eurostat.

Dominitz J., Sherman R.P. (2006), Identification and estimation of bounds on school performance measures: a nonparametrci analysis of a mixture model with verification, *Journal of Applied Econometrics*, 21: 1295–1326.

Epland J., Kirkeberg M.I. (2002), Comparing Norwegian income data in administrative registers with income data in the Survey of Living Conditions, paper presented at The International Conference on Improving Surveys (ICIS), Copenhagen, Denmark, August 25–28, 2002.

Groves R.M. (1989), *Survey Errors and Survey Costs*, Wiley, New York.

Groves R.M., Dillman D.A., Eltinge J.L., Little R.J.A. (2002), *Survey Nonresponse*, Wiley, New York.

Guio A.C. (2004), Income poverty and social exclusion in the EU25, *Population and Social Condition*, 13/2005, Eurostat.

Horowitz J.L., Manski C.F. (1995), Identification and robustness with contaminated and corrupted data, *Econometrica*, 63: 281-302.

Horowitz J.L., Manski C.F. (1998), Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputation, *Journal of Econometrics*, 84: 37–58.

Kreider B., Pepper J.V. (2007), Disability and employment: reevaluating the evidence in light of reporting errors, *Journal of the American Statistical Association*, 101: 432–441.

Lessler J.T., Kalsbeek W.D. (1992), *Nonsampling Error in Surveys*, Wiley, New York.

Little J.A., Rubin D.B. (1987), *Statistical Analysis with Missing Data*, Wiley, New York.

Manski C.F. (1989), Anatomy of the selection problem, *Journal of Human Resources*, 24: 343–360.

Manski C.F. (2003), *Partial Identification of Probability Distributions*, Springer, New York.

Manski C.F., Pepper J. (2000), Monotone instrumental variables with an application to the returns to schooling, *Econometrica*, 68: 997–1010.

Manski C.F., Tamer E. (2002), Inference on regressions with interval data on a regressor or outcome, *Econometrica*, 70: 519–546.

Molinari F. (2005), Partial identification of probability distributions with misclassified data, mimeo, Cornell University.

Nicoletti C. (2003), Poverty analysis with item and unit nonresponses: Alternative estimators compared, ISER Working Paper 2003-20, University of Essex.

Nicoletti C., Peracchi F. (2002), A cross-country comparison of survey participation in the ECHP, ISER Working Paper 2002-32, University of Essex.

Pudney S., Francavilla F. (2006), Income mis-measurement and the estimation of poverty rates. An Analysis of income poverty in Albania, ISER Working Paper 2006-35, University of Essex.

Ravallion M. (1994), Poverty rankings using noisy data on living standards, *Economics Letters*, 45: 481–485.

Rubin D.B. (1989), *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Rubin D.B. (1996), Multiple imputation after 18+ years, *Journal of the American Statistical Association*, 91: 473–520.

van Praag B., Hagenaars A., van Eck W. (1983), The influence of classification and observation errors on the measurement of income inequality, *Econometrica*, 51: 1093- 1108.

Vasquez-Alvarez V.R., Melenberg B., van Soest A. (1999), Bounds on quantiles in the presence of full and item nonresponse, CentER Discussion Paper 1999–38, Tilburg University.

Vasquez-Alvarez V.R., Melenberg B., van Soest A. (2001), Nonparametric bounds in the presence of item nonrespone, unfolding brackets, and anchoring, CentER Discussion Paper 2001–67, Tilburg University.

Vella F. (1998), Estimating models with sample selection bias: A survey, *Journal of Human Resources*, 33: 127-169.

Table 1: Point estimates of poverty rates by country in 2001 (standard errors in parentheses). For each country, the first row reports the estimates of the poverty rates, while the second row reports the corresponding standard errors in parenthesis.

Country	No. obs.	Nonresponse rate	Unweighted poverty	Weighted poverty
Belgium	5607	.201	.131	.115
			(.005)	(.004)
Denmark	4975	.144	.094	.087
			(.004)	(.004)
France	12625	.100	.180	.184
			(.003)	(.003)
Germany (SOEP)	13489	.157	.132	.136
			(.003)	(.003)
Greece	11114	.131	.205	.253
			(.004)	(.004)
Ireland	5421	.099	.175	.160
			(.005)	(.005)
Italy	15317	.190	.223	.241
			(.003)	(.003)
Netherlands	10395	.073	.139	.147
			(.003)	(.003)
Portugal	12917	.138	.202	.241
			(.004)	(.004)
Spain	13689	.123	.185	.193
			(.003)	(.003)
UK (BHPS)	10805	.102	.187	.174
			(.004)	(.004)

Table 2: Estimated bounds by country ($\lambda = 0.100$, $\rho = 0.025$). For each country, the estimates of the upper (lower) bounds are reported in the first row, while the corresponding upper (lower) limits of the bootstrap confidence intervals are reported in the second row.

Country	LB_{C1^*}	UB_{C1^*}	LB_{C2^*}	UB_{C2^*}	LB_{C3^*}	UB_{C3^*}
Belgium	.035	.258	.103	.246	.039	.191
	.018	.295	.086	.284	.021	.228
Denmark	.000	.221	.074	.213	.000	.144
	.000	.285	.052	.277	.000	.209
France	.067	.270	.141	.254	.074	.197
	.055	.288	.130	.273	.061	.216
Germany	.020	.250	.094	.240	.022	.177
	.009	.279	.083	.269	.010	.206
Greece	.099	.342	.168	.324	.110	.275
	.086	.368	.155	.350	.095	.302
Ireland	.083	.327	.156	.310	.093	.256
	.056	.386	.129	.370	.062	.316
Italy	.109	.344	.172	.325	.121	.283
	.094	.373	.157	.355	.104	.313
Netherlands	.046	.278	.125	.264	.052	.200
	.034	.306	.113	.293	.038	.229
Portugal	.106	.351	.170	.332	.118	.289
	.081	.415	.145	.397	.090	.353
Spain	.074	.316	.143	.300	.082	.249
	.060	.355	.128	.339	.067	.287
UK	.072	.322	.147	.305	.080	.249
	.061	.347	.135	.331	.067	.275

Table 3: Estimates of UB_D, LB_D and of the width $\mathrm{UB}_D - \mathrm{LB}_D$ by country ($\lambda = 0.100$, $\rho = 0.025$). For each country, the estimates of the upper (lower) bounds are reported in the first row, while the corresponding upper (lower) limits of the bootstrap confidence intervals are reported in the second row. W_1 is the part of the interval width due to missing data problems, while W_2 is that due to measurement error problems.

Country	LB_D	UB_D	Width	W_1	%	W_2	%
Belgium	.103	.191	.089	.064	72.792	.024	27.208
	.086	.228	.142	.079	55.412	.063	44.588
Denmark	.074	.144	.071	.054	75.670	.017	24.330
	.052	.209	.157	.084	53.411	.073	46.589
France	.141	.197	.055	.022	40.043	.033	59.957
	.130	.216	.086	.027	31.152	.059	68.848
Germany (SOEP)	.094	.177	.084	.062	73.745	.022	26.255
	.083	.206	.123	.075	60.854	.048	39.146
Greece	.168	.275	.108	.068	63.346	.039	36.654
	.155	.302	.146	.078	53.504	.068	46.496
Ireland	.156	.256	.100	.063	63.370	.037	36.630
	.129	.316	.187	.087	46.551	.100	53.449
Italy	.172	.283	.111	.071	63.741	.040	36.259
	.157	.313	.156	.082	52.289	.075	47.711
Netherlands	.125	.200	.075	.046	60.780	.029	39.220
	.113	.229	.116	.057	48.851	.059	51.149
Portugal	.170	.289	.118	.078	66.169	.040	33.831
	.145	.353	.209	.108	51.670	.101	48.330
Spain	.143	.249	.106	.072	68.359	.033	31.641
	.128	.287	.159	.090	56.346	.070	43.654
UK (BHPS)	.147	.249	.103	.068	66.510	.034	33.490
	.135	.275	.140	.078	56.036	.061	43.964

Table 4: Minimum and maximum widths across countries of the estimated interval $[\mathrm{LB}_D, \mathrm{UB}_D]$ for different values of λ and $\rho = 0.25\lambda$.

λ	ρ	min width	max width	min W1 %	max W1 %
.010	.003	.024	.080	.866	.969
.050	.013	.037	.096	.568	.862
.100	.025	.055	.118	.400	.757
.200	.050	.093	.166	.254	.603
.300	.075	.118	.221	.185	.495
.500	.125	.185	.371	.112	.339
.750	.188	.375	.720	.055	.183
.990	.248	.882	.972	.042	.110

Table 5: Minimum and maximum widths across countries of the estimated interval $[\mathrm{LB}_D, \mathrm{UB}_D]$ for $\lambda = 0.100$ and different values of ρ.

ρ	min width	max width	min W1 %	max W1 %
.025	.055	.118	.400	.757
.050	.057	.120	.420	.764
.100	.061	.125	.456	.778
.200	.069	.134	.516	.802
.300	.076	.143	.564	.821
.500	.087	.165	.636	.850
.750	.093	.195	.685	.876
.990	.099	.230	.704	.895

A The extended indirect misclassification model

The bounds obtained in Section 2.3 may also be derived by considering an extended version of the indirect misclassification model (5). The main advantage of this approach is that it can be used to identify bounds on the distribution of any categorical variable (not just a binary indicator) in the presence of missing or interval data and of measurement error problems.

The bounds (8)–(10) can be derived by assuming either C1, C2 or C3, and then considering the following extended indirect misclassification model

$$\begin{bmatrix} \Pr(D_Y = 1) \\ \Pr(D_Y = 0) \end{bmatrix} = P \begin{bmatrix} \Pr(D_W = 1, D_R = 1) \\ \Pr(D_W = 0, D_R = 1) \\ \Pr(D_R = 0) \end{bmatrix}, \tag{15}$$

where P is a rectangular matrix given by

$$P = \begin{bmatrix} \Pr(D_Y = 1 \,|\, D_W = 1, D_R = 1) & \Pr(D_Y = 1 \,|\, D_W = 0, D_R = 1) & \Pr(D_Y = 1 \,|\, D_R = 0) \\ \Pr(D_Y = 0 \,|\, D_W = 1, D_R = 1) & \Pr(D_Y = 0 \,|\, D_W = 0, D_R = 1) & \Pr(D_Y = 0 \,|\, D_R = 0) \end{bmatrix}.$$

This model extends the indirect classification model (5) by adding the probability $\Pr(D_R = 0)$ of missing data. Model (15) can be split into the sum of two terms

$$\begin{bmatrix} \Pr(D_Y = 1) \\ \Pr(D_Y = 0) \end{bmatrix} = \begin{bmatrix} \Pr(D_Y = 1 \,|\, D_R = 1) \\ \Pr(D_Y = 0 \,|\, D_R = 1) \end{bmatrix} \Pr(D_R = 1) + \begin{bmatrix} \Pr(D_Y = 1 \,|\, D_R = 0) \\ \Pr(D_Y = 0 \,|\, D_R = 0) \end{bmatrix} \Pr(D_R = 0), \tag{16}$$

where

$$\begin{bmatrix} \Pr(D_Y = 1 \,|\, D_R = 1) \\ \Pr(D_Y = 0 \,|\, D_R = 1) \end{bmatrix} = \begin{bmatrix} \Pr(D_Y = 1 \,|\, D_W = 1, D_R = 1) & \Pr(D_Y = 1 \,|\, D_W = 0, D_R = 1) \\ \Pr(D_Y = 0 \,|\, D_W = 1, D_R = 1) & \Pr(D_Y = 0 \,|\, D_W = 0, D_R = 1) \end{bmatrix} \times \begin{bmatrix} \Pr(D_W = 1 \,|\, D_R = 1) \\ \Pr(D_W = 0 \,|\, D_R = 1) \end{bmatrix} \tag{17}$$

is just the indirect misclassification model introduced in Section 2.1 except for the fact that all probabilities are now conditional on $D_R = 1$. Given assumption C1, C2 or C3 and splitting the extended misclassification model in two additional terms as in (16), we can identify an upper (lower) bound for the probability of being poor by computing an upper (lower) bound for the probability of being poor separately for respondents (under Assumptions C1, C2 or C3) and non-respondents.

The bounds (12)–(14) can be derived by imposing Assumption C4 plus one of the Assumptions C1, C2 or C3, and by considering the following extended indirect misclassification model

$$\begin{bmatrix} \Pr(D_Y = 1) \\ \Pr(D_Y = 0) \end{bmatrix} = P \begin{bmatrix} \Pr(D_W = 1, D_R = 1) \\ \Pr(D_W = 0, D_R = 1) \\ \Pr(D_{W^*} = 1, D_R = 0) \\ \Pr(D_{W^*} = 0, D_R = 0) \end{bmatrix},$$

where P is a rectangular matrix whose transpose is equal to

$$P' = \begin{bmatrix} \Pr(D_Y = 1 \,|\, D_W = 1, D_R = 1) & \Pr(D_Y = 0 \,|\, D_W = 1, D_R = 1) \\ \Pr(D_Y = 1 \,|\, D_W = 0, D_R = 1) & \Pr(D_Y = 0 \,|\, D_W = 0, D_R = 1) \\ \Pr(D_Y = 1 \,|\, D_{W^*} = 1, D_R = 0) & \Pr(D_Y = 0 \,|\, D_{W^*} = 1, D_R = 0) \\ \Pr(D_Y = 1 \,|\, D_{W^*} = 0, D_R = 0) & \Pr(D_Y = 0 \,|\, D_{W^*} = 0, D_R = 0) \end{bmatrix}.$$

It is again possible to split this extended model into the sum of two parts as in (16) where (17) still holds and

$$\begin{bmatrix} \Pr(D_Y = 1 \,|\, D_R = 0) \\ \Pr(D_Y = 0 \,|\, D_R = 0) \end{bmatrix} = \begin{bmatrix} \Pr(D_Y = 1 \,|\, D_{W^*} = 1, D_R = 0) & \Pr(D_Y = 1 \,|\, D_{W^*} = 0, D_R = 0) \\ \Pr(D_Y = 0 \,|\, D_{W^*} = 1, D_R = 0) & \Pr(D_Y = 0 \,|\, D_{W^*} = 0, D_R = 0) \end{bmatrix} \times \begin{bmatrix} \Pr(D_{W^*} = 1 \,|\, D_R = 0) \\ \Pr(D_{W^*} = 0 \,|\, D_R = 0) \end{bmatrix}, \tag{18}$$

with $\Pr(D_Y = 1 \,|\, D_R = 0)$ defined as in (11) in Section 2.3.

Notice that using the matrix notation it is straightforward to extend the indirect misclassification model to the case of interval data. Let us assume, for example, that unfolding bracket questions are used to collect some partial information on income for non-respondents. Then we can observe an upper and a lower bound for the household income Y_U and Y_L. Notice that if respondents refuse to give any information on household income, then Y_U is infinite and Y_L is zero. The new indirect misclassification model is given by

$$\begin{bmatrix} \Pr(D_Y = 1) \\ \Pr(D_Y = 0) \end{bmatrix} = P \begin{bmatrix} \Pr(D_W = 1, D_R = 1) \\ \Pr(D_W = 0, D_R = 1) \\ \Pr(D_{Y_L} = 1, D_{Y_U} = 1, D_R = 0) \\ \Pr(D_{Y_L} = 0, D_{Y_U} = 1, D_R = 0) \\ \Pr(D_{Y_L} = 0, D_{Y_U} = 0, D_R = 0) \end{bmatrix},$$

where D_{Y_L} and D_{Y_U} are binary indicators equal to one if, respectively, $Y_L \leq \gamma$ and $Y_U \leq \gamma$, and equal to zero otherwise, and P is a rectangular matrix whose transpose is equal to

$$P' = \begin{bmatrix} \Pr(D_Y = 1 \,|\, D_W = 1, D_R = 1) & \Pr(D_Y = 0 \,|\, D_W = 1, D_R = 1) \\ \Pr(D_Y = 1 \,|\, D_W = 0, D_R = 1) & \Pr(D_Y = 0 \,|\, D_W = 0, D_R = 1) \\ \Pr(D_Y = 1 \,|\, D_{Y_L} = 1, D_{Y_U} = 1, D_R = 0) & \Pr(D_Y = 0 \,|\, D_{Y_L} = 1, D_{Y_U} = 1, D_R = 0) \\ \Pr(D_Y = 1 \,|\, D_{Y_L} = 0, D_{Y_U} = 1, D_R = 0) & \Pr(D_Y = 0 \,|\, D_{Y_L} = 0, D_{Y_U} = 1, D_R = 0) \\ \Pr(D_Y = 1 \,|\, D_{Y_L} = 0, D_{Y_U} = 0, D_R = 0) & \Pr(D_Y = 0 \,|\, D_{Y_L} = 0, D_{Y_U} = 0, D_R = 0) \end{bmatrix}.$$

As before it is possible to split this extended model into the sum of two parts as in (16) where (17)

still holds and

$$\begin{bmatrix} \Pr(D_Y = 1 \,|\, D_R = 0) \\ \Pr(D_Y = 0 \,|\, D_R = 0) \end{bmatrix} = \\ \begin{bmatrix} \Pr(D_Y = 1 \,|\, D_{Y_L} = 1, D_{Y_U} = 1, D_R = 0) & \Pr(D_Y = 0 \,|\, D_{Y_L} = 1, D_{Y_U} = 1, D_R = 0) \\ \Pr(D_Y = 1 \,|\, D_{Y_L} = 0, D_{Y_U} = 1, D_R = 0) & \Pr(D_Y = 0 \,|\, D_{Y_L} = 0, D_{Y_U} = 1, D_R = 0) \\ \Pr(D_Y = 1 \,|\, D_{Y_L} = 0, D_{Y_U} = 0, D_R = 0) & \Pr(D_Y = 0 \,|\, D_{Y_L} = 0, D_{Y_U} = 0, D_R = 0) \end{bmatrix}' \\ \times \begin{bmatrix} \Pr(D_{Y_L} = 1, D_{Y_U} = 1 \,|\, D_R = 0) \\ \Pr(D_{Y_L} = 0, D_{Y_U} = 1 \,|\, D_R = 0) \\ \Pr(D_{Y_L} = 0, D_{Y_U} = 0 \,|\, D_R = 0) \end{bmatrix}. \tag{19}$$

The probabilities in the last vector can be computed using the interval information on income. Moreover, we know that

$$\Pr(D_Y = 1 \,|\, D_{Y_L} = 0, D_{Y_U} = 0, D_R = 0) = 0, \quad \Pr(D_Y = 0 \,|\, D_{Y_L} = 0, D_{Y_U} = 0, D_R = 0) = 1, \\ \Pr(D_Y = 1 \,|\, D_{Y_L} = 1, D_{Y_U} = 1, D_R = 0) = 1, \quad \Pr(D_Y = 0 \,|\, D_{Y_L} = 1, D_{Y_U} = 1, D_R = 0) = 0. \tag{20}$$

Bounding the remaining unknown probabilities in (19) in the closed interval $[0, 1]$, we can easily compute upper and lower bounds for $\Pr(D_Y = 1 \,|\, D_R = 0)$ and its compliment to one.

If Y_U, Y_L and the poverty line are measured with errors, then the equalities in (20) could be invalid. In that case we need to introduce the following additional assumptions, analogous to assumption C1 in Section 2.3:

$$\Pr(D_Y = 1 \,|\, D_{Y_L} = 0, D_{Y_U} = 0, D_R = 0) \leq \delta_1 < 1. \\ \Pr(D_Y = 0 \,|\, D_{Y_L} = 1, D_{Y_U} = 1, D_R = 0) \leq \delta_2 < 1,$$

where δ_1 and δ_2 are fixed constants.

Finally, notice that the extended indirect misclassification model can be applied to the case where all individuals provided only interval information on their household income. In this case the model would coincide with equation (19) but without conditioning on $D_R = 0$.

B Proofs

Proof of Proposition 1 We begin by showing that Assumption B1 implies the bounds UB_{B1} and LB_{B1} on $\Pr(D_Y = 1)$. By the law of total probability, the true poverty probability can be expressed as

$$\Pr(D_Y = 1) = \Pr(D_Y = 1 \,|\, D_W = 1) \Pr(D_W = 1) + \Pr(D_Y = 1 \,|\, D_W = 0) \Pr(D_W = 0). \tag{21}$$

Since $\Pr(D_Y = 1 \,|\, D_W = 1) = 1 - \Pr(D_Y = 0 \,|\, D_W = 1)$, we can rewrite (21) as

$$\Pr(D_Y = 1) = \Pr(D_W = 1) + \Pr(D_Y = 1 \,|\, D_W = 0) \Pr(D_W = 0)$$
$$- \Pr(D_Y = 0 \,|\, D_W = 1) \Pr(D_W = 1).$$

By Assumption B1, $0 \leq \Pr(D_W = i, D_Y = j) \leq \lambda < 1$ for any $i \neq j$ and we can identify the following upper and lower bound for $\Pr(D_Y = 1)$

$$\text{UB}_{B1} = \min(\Pr(D_W = 1) + \lambda, 1),$$
$$\text{LB}_{B1} = \max(\Pr(D_W = 1) - \lambda, 0).$$

The same upper and lower bounds can be obtained by considering the direct misclassification model (3)–4) and Assumption B1 as proved in Molinari (2005), Proposition 3.

We now show that Assumption B2 implies the bounds UB_{B2} and LB_{B2} on $\Pr(D_Y = 1)$. Consider again equation (21). Using the fact that Assumption B2 implies that $\Pr(D_Y = 1 \,|\, D_W = 0) \leq \lambda$ and $\Pr(D_Y = 1 \,|\, D_W = 1) \geq 1 - \lambda$, we obtain the following upper and lower bounds

$$\text{UB}_{B2} = \Pr(D_W = 1) + \lambda \Pr(D_W = 0),$$
$$\text{LB}_{B2} = \Pr(D_W = 1)(1 - \lambda).$$

It is straightforward to show that the above bounds are sharp. For example, to show that UB_{B2} is a sharp upper bound under Assumption B2, we just need to show that there is no $\epsilon < 0$ such that

$$UB_{B2}(\epsilon) = \Pr(D_W = 1) + \lambda \Pr(D_W = 0) + \epsilon$$

is a valid bound for $\Pr(D_Y = 1)$ satisfying

$$\Pr(D_Y = 1) = \Pr(D_Y = 1 \,|\, D_W = 1) \Pr(D_W = 1) + \Pr(D_Y = 1 \,|\, D_W = 0) \Pr(D_W = 0)$$

for any value of $\Pr(D_Y = 1 \,|\, D_W = 0)$ in $[0, \lambda]$ and any value of $\Pr(D_Y = 1 \,|\, D_W = 1)$ in $[1 - \lambda, 1]$. Since for $\Pr(D_Y = 1 \,|\, D_W = 0) = \lambda$ and $\Pr(D_Y = 1 \,|\, D_W = 1) = 1$ we have

$$\Pr(D_Y = 1) = \Pr(D_W = 1) + \lambda \Pr(D_W = 0),$$

$UB_{B2}(\epsilon)$ is a valid upper bound if and only if $\epsilon \geq 0$, which implies that UB_{B2} is sharp. The proof that the remaining bounds (UB_{B1}, LB_{B1} and LB_{B2}) are sharp can be obtained in a similar way.

Proof of Proposition 2 We begin by showing that Assumption C1 implies the bounds UB_{C1} and LB_{C1} on $\Pr(D_Y = 1)$. By the law of total probability, the poverty rate satisfies

$$\Pr(D_Y = 1) = \Pr(D_Y = 1 \,|\, D_R = 1)\ \Pr(D_R = 1) + \Pr(D_Y = 1 \,|\, D_R = 0)\ \Pr(D_R = 0). \tag{22}$$

By conditioning on the event $D_R = 1$ and applying Proposition 1, it is possible to prove that Assumption C1 implies

$$\max\{\Pr(D_W = 1 \,|\, D_R = 1) - \lambda, 0\} \leq \Pr(D_Y = 1 \,|\, D_R = 1) \leq \min\{\Pr(D_W = 1 \,|\, D_R = 1) + \lambda, 1\}.$$

By replacing these bounds in equation (22) and bounding $\Pr(D_Y = 1 \,|\, D_R = 0)$ in the closed interval $[0, 1]$ we can derive the following upper and lower bounds for $\Pr(D_Y = 1)$

$$\text{UB}_{C1} = \min\{\Pr(D_W = 1 \,|\, D_R = 1) + \lambda, 1\}\ \Pr(D_R = 1) + \Pr(D_R = 0),$$
$$\text{LB}_{C1} = \max\{\Pr(D_W = 1 \,|\, D_R = 1) - \lambda, 0\}\ \Pr(D_R = 1).$$

The proof that Assumption C2 implies the bounds UB_{C2} and LB_{C2} on $\Pr(D_Y = 1)$ can be obtained by analogy with the previous case.

The proof that Assumption C3 implies the bounds UB_{C3} and LB_{C3} on $\Pr(D_Y = 1)$ follows the same lines except for the fact that bounds on $\Pr(D_Y = 1 \,|\, D_R = 1)$ are derived using Assumption C3 and Proposition 3 in Molinari (2005). □

NON-TECHNICAL SUMMARY

The aim of this paper is estimating income poverty in the presence of missing data and measurement error problems. We use the conventional definition of poverty as having an income below a poverty line, 60% of median household income, where income is scaled to take account of household composition and size.

Income measures are usually among the variables with the highest nonresponse rates. Both survey methodologists and applied econometricians suggest that ignoring the nonresponse problem can cause a serious selection bias. Moreover, since responses to surveys are not perfectly reliable, income poverty measures are generally plagued by measurement errors too.

Point estimation approaches taking account of missing data and/or measurement error problems impose usually restrictive and non-testable assumptions. On the contrary, in this paper we do not impose any restrictive assumption but we provide bound estimates instead of point estimates of the poverty rate. In other words, we provide an upper and a lower bound which defines the range of logically possible values for the poverty rate.

By using the European Community Household Panel (ECHP) we compute bound estimates for the poverty rates in Belgium, Denmark, France, Germany, Greece, Ireland, Italy, the Netherlands, Portugal, Spain, and the UK. Interestingly, our bound estimates are in line with the official statistics published by Eurostat, where Greece, Italy and Portugal are the countries with the highest poverty rates, around 23 percent, while Belgium, Denmark, Germany and the Netherlands have the lowest poverty rates, usually lower than 15 percent. In fact, our estimated lower bounds for the poverty rate are higher than 15 percent in Greece, Italy and Portugal, while the estimated upper bounds are lower than 23 percent in Belgium, Denmark, Germany and the Netherlands. Moreover, we find that the missing data problem is more relevant than the measurement error problem in all countries.