

Nicoletti, Cheti; Peracchi, Franco; Foliano, Francesca

**Working Paper**

## Estimating Income Poverty in the Presence of Missing Data and Measurement Error

SOEPpapers on Multidisciplinary Panel Data Research, No. 252

**Provided in Cooperation with:**

German Institute for Economic Research (DIW Berlin)

*Suggested Citation:* Nicoletti, Cheti; Peracchi, Franco; Foliano, Francesca (2009) : Estimating Income Poverty in the Presence of Missing Data and Measurement Error, SOEPpapers on Multidisciplinary Panel Data Research, No. 252, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:

<https://hdl.handle.net/10419/150798>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

# SOEPpapers

on Multidisciplinary Panel Data Research

# 252

Cheti Nicoletti • Franco Peracchi • Francesca Foliano

## Estimating Income Poverty in the Presence of Missing Data and Measurement Error

Berlin, December 2009

## **SOEPpapers on Multidisciplinary Panel Data Research** at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel Study (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPpapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

Any opinions expressed in this series are those of the author(s) and not those of DIW Berlin. Research disseminated by DIW Berlin may include views on public policy issues, but the institute itself takes no institutional policy positions.

The SOEPpapers are available at  
**<http://www.diw.de/soeppapers>**

### **Editors:**

Georg **Meran** (Dean DIW Graduate Center)

Gert G. **Wagner** (Social Sciences)

Joachim R. **Frick** (Empirical Economics)

Jürgen **Schupp** (Sociology)

Conchita **D'Ambrosio** (Public Economics)

Christoph **Breuer** (Sport Science, DIW Research Professor)

Anita I. **Drever** (Geography)

Elke **Holst** (Gender Studies)

Martin **Kroh** (Political Science and Survey Methodology)

Frieder R. **Lang** (Psychology, DIW Research Professor)

Jörg-Peter **Schräpler** (Survey Methodology)

C. Katharina **Spieß** (Educational Science)

Martin **Spieß** (Survey Methodology, DIW Research Professor)

ISSN: 1864-6689 (online)

German Socio-Economic Panel Study (SOEP)  
DIW Berlin  
Mohrenstrasse 58  
10117 Berlin, Germany

Contact: Uta Rahmann | [urahmann@diw.de](mailto:urahmann@diw.de)

# Estimating Income Poverty in the Presence of Missing Data and Measurement Error

Cheti Nicoletti  
ISER, University of Essex  
Colchester, UK

Franco Peracchi  
Tor Vergata University and EIEF  
Rome, Italy

Francesca Foliano  
Tor Vergata University  
Rome, Italy

## Abstract

Reliable measures of poverty are an essential statistical tool for 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 missing data and measurement error. Neglecting these problems can bias the estimated poverty rates. We show how to derive upper and lower bounds for the population poverty rate using the sample evidence, an upper bound on the probability of misclassifying people into poor and non-poor, and instrumental or monotone instrumental variable assumptions. By using the European Community Household Panel, we compute bounds for the poverty rate in ten European countries and study the sensitivity of poverty comparisons across countries to missing data and measurement error problems. Supplemental materials for this article may be downloaded from the *JBES* website.

**KEY WORDS:** Misclassification error; Survey non-response; Partial identification.

# 1 INTRODUCTION

Income poverty measures are designed to count the poor and to diagnose the extent and distribution of poverty. For this reason, they are an essential statistical tool for public policies aimed at reducing poverty (Deaton 1997). Estimation of income poverty is usually based on survey data and is typically plagued by missing data and measurement error.

Missing data arises from the failure to obtain a complete response from all individuals included in a survey. It may occur because individuals refuse to return their questionnaire (unit nonresponse) or do not provide an answer for some of the questions (item nonresponse), and may depend on both individual attitudes and survey procedures. Measurement error represents instead the deviation between the recorded answer to a survey question and the underlying attribute being measured. It may reflect systematic misreporting or unreliable response by the interviewee, and may depend on data collection procedures (questionnaire design and interview methods), the way the interviewer interacts with the interviewee, and data processing (data entry, editing, coding, etc).

Missing data and measurement error are especially important in the case of income. Questions about income are sensitive in nature and people may refuse to answer because of privacy invasion or a perceived risk of disclosure of information to third parties. Moreover, even when people are willing to report their income, they might misreport it because of memory problems or a tendency to overestimate or underestimate it.

Imputation and weighting methods are the approaches to missing data usually adopted by survey methodologists (see Little and Rubin 1987, and Rubin 1989, 1996). They typically assume a missing at random (MAR) condition, that is, 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 also take into account selection due to unobserved variables (see Vella 1998 for a survey). While these methods relax the MAR condition, they typically impose various types of restrictions on the distribution of the unobservables.

The most common statistical approaches to measurement error 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 general

survey of the literature). 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 violation of this assumption occurs 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 for a fraction of individuals and that the observed outcome is equal to a mixture of two variables, the true outcome and an unknown contaminating variable.

Most estimation methods proposed for missing data or measurement error 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 additional assumptions, to learn something about the parameters of interest. Their approach involves a shift from point identification to partial identification, 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 and the maintained assumptions.

In this paper we follow the partial identification approach and provide bounds on poverty rate in ten European countries using the microdata from the last wave (2001) of the European Community Household Panel. These bounds take account of the presence of both measurement error and missing data problems, and are meant to establish a “domain of consensus” that represents a starting point for subsequent analyses. To our knowledge, this is the first study which formally considers identification issues caused by the presence of both types of problems. We combine results in Nicoletti and Peracchi (2002) and Nicoletti (2008) to bound the poverty rate in the presence of missing data with the approach suggested by Horowitz and Manski (1995) and Molinari (2003, 2008) to take measurement errors into account.

The data used in our application are described in Section 2. We formalize the partial identification approach to poverty rates in Section 3, first in the presence of either missing data or measurement errors, and then in the presence of both together. We derive analytical bounds by

exploiting the availability of partial information on income under different assumptions on the probability of misclassifying poverty status. Section 4 presents our empirical results. Finally, Section 5 draws some conclusions.

## 2 DATA

We begin by describing the problems that arise when estimating poverty measures using the European Community Household Panel (ECHP), a dataset that we take as representative of the kind of survey data typically used for this purpose.

The ECHP is a longitudinal household survey centrally designed and coordinated by the Statistical Office of the European Communities (Eurostat) and conducted annually from 1994 to 2001. The ECHP is patterned after the U.S. Panel Study of Income Dynamics, and was explicitly designed to derive indicators of poverty and social exclusion for the European Commission. Its target population consists of all individuals living in private households in the 15 member countries of the European Union before its enlargement. All sampled individuals aged 16 or more are asked to complete a personal questionnaire. Moreover, a reference person in each household, usually the household head or the spouse/partner of the head, is asked to fill-in a household questionnaire.

In its first wave (1994), the survey covered about 60,000 households and 130,000 individuals in 12 countries, namely Belgium, Denmark, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain and the UK. Austria, Finland and Sweden began to participate in the ECHP only later, respectively from the second (1995), third (1996) and fourth (1997) wave. We exclude the countries which did not participate for the whole period 1994–2001. We also exclude France because of the doubtful quality of the gross/net conversion of income variables. This gives a sample of 10 countries, namely Belgium, Denmark, Germany, Greece, Ireland, Italy, the Netherlands, Portugal, Spain and the UK.

We focus on nonresponse and measurement error on household income for individuals belonging to responding households, namely those for which at least the household questionnaire was returned. The resulting sample consists of the 103,605 individuals observed in the most recent wave (2001). In all our empirical applications, we take account of sampling design and the presence of nonresponding households (those for which no questionnaire was returned) by using the weights provided in the

public-use files of the ECHP.

Our poverty measure is the headcount ratio, namely the fraction of people (both children and adults) living in households with income below a certain threshold (the “poverty line”). For brevity, we refer to this measure as the poverty rate. The key variable in the construction of our poverty measure is total net household income, computed in the ECHP as the sum of all annual incomes (wages and salaries, self-employment income, pensions, etc.) reported by all members of a household. Annual income is the amount received in the year before the survey, net of taxes and expressed in national currency and current prices. Following conventional practice, we divide real household income by the modified OECD equivalence scale to take account of household size and composition. We define an individual as poor if her equivalized household income is below a poverty line defined as 60% of the national median value, estimated using the imputed values and the sampling weights provided by the ECHP.

Because of the way household income is constructed, nonresponse may occur either because of item nonresponse to some income questions or because of unit nonresponse by household members who fail to return the personal questionnaire. While income nonresponse can be observed (see the last column of Table 1), the amount of measurement error cannot. The assessment of measurement error requires validation studies (see for example Bound and Krueger 1991, Rodgers et al. 1993, and Bound et al. 1994). In this paper we focus on misclassification error, namely measurement error in the indicator of poverty status. A useful source of information in this case is the validation study of Epland and Kirkeberg (2002), who compare true and reported income by matching administrative data with the 1996 Norwegian Survey of Living Conditions. We use their results in our empirical application to impose credible assumptions on misclassification probabilities.

Table 1 shows, for each of the countries considered, point estimates of the population poverty rates and their estimated standard errors (in parenthesis). We report estimates computed ignoring individuals with nonresponse to household income (poverty rates for respondents) and estimates that use the imputed income values provided by the ECHP (poverty rates with imputation).

Ignoring income nonresponse does not cause any bias when data are missing completely at random (MCAR), that is, when the response probability is constant across individuals. This assumption contrasts sharply with the evidence from the ECHP, where nonresponse can be predicted



using variables such as household size, the number of active members in the household, the level of education of the household head, and characteristics of the data collection process (Nicoletti and Peracchi 2002). Using imputed values to replace missing income is the standard approach adopted to compute poverty rates in official statistics. This produces unbiased estimates of poverty rates only if the data is missing at random (MAR) and the imputation model is correct. Since MAR is an untestable assumption, however, it is impossible to evaluate the potential bias caused by imputation.

In the rest of this paper we check whether relaxing these untestable assumptions still allows us to identify meaningful bounds on the population poverty rates. As we argue in the next section, the fraction of income nonrespondents and the probability of misclassifying poverty status are a direct measure of how severe the identification problem is. Since nonresponse rates and misclassification probabilities are usually not small, the identified bounds can be wide. For this reason, in the next section we suggest to narrow the bounds by using partial information on income, by introducing some untestable but “credible” assumptions on the misclassification process and by imposing some instrumental and monotone instrumental variable assumptions.

### 3 PARTIAL IDENTIFICATION OF POVERTY RATES

We consider partial identification of population poverty rates from data subject to nonsampling errors similar to those that plague the ECHP. Section 3.1 considers the case of missing data but no measurement error, Section 3.2 considers the case of measurement error but no missing data problems, while Section 3.3 considers the case of both missing data and measurement error.

#### 3.1 Partial Identification in the Presence of Missing Data

Let  $Y$  represent the equivalized income of a household, let  $\gamma$  be the poverty line, and let  $D_Y$  be the indicator of poverty status, equal to one if a person lives in a household with  $Y \leq \gamma$  and equal to zero otherwise. The population poverty rate is the fraction of people living in households for which  $Y$  does not exceed  $\gamma$ . Formally, the population poverty rate is just  $\Pr(D_Y = 1) = \Pr(Y \leq \gamma)$ .

Suppose that there is no measurement error in  $Y$  and  $\gamma$  but, because of nonresponse, household income is missing for a fraction of the individuals. Following Manski (1989), let  $D_R$  be a binary

indicator equal to one if a person belongs to a responding household, namely one whose income is fully reported, and equal to zero otherwise. By the law of total probability, the population poverty rate satisfies

$$\Pr(D_Y = 1) = \Pr(D_Y = 1 \mid D_R = 1) \Pr(D_R = 1) + \Pr(D_Y = 1 \mid D_R = 0) \Pr(D_R = 0). \quad (1)$$

Because only  $\Pr(D_Y = 1 \mid D_R = 1)$ ,  $\Pr(D_R = 1)$  and  $\Pr(D_R = 0)$  can be point-identified from the sampling process, the population poverty rate is not point-identified unless additional assumptions are made. However, it is partially identified by the fact that the unknown element  $\Pr(D_Y = 1 \mid D_R = 0)$  must necessarily lie between zero and one. Substituting these bounds in (1) gives the following upper and lower bounds on the population poverty rate

$$\text{UB} = \Pr(D_Y = 1 \mid D_R = 1) \Pr(D_R = 1) + \Pr(D_R = 0),$$

$$\text{LB} = \Pr(D_Y = 1 \mid D_R = 1) \Pr(D_R = 1).$$

These bounds are sharp, that is, they exhaust the information about  $\Pr(D_Y = 1)$  available from the sampling process and the maintained assumptions. The width  $\text{UB} - \text{LB}$  of the identification region for  $\Pr(D_Y = 1)$  is equal to the nonresponse probability  $\Pr(D_R = 0)$ , which therefore represents a direct measure of the uncertainty about the population poverty rate caused by nonresponse.

An important question is how to shrink these “worst-case” bounds, that is, how to narrow the identification region for the population poverty rate. One possibility is to impose instrumental variable (IV) restrictions. A random variable  $Z$ , with values in a subset  $\mathcal{Z}$  of the real line, is an IV if it helps predict response but does not help predict poverty, possibly after conditioning on a set  $X$  of observable covariates with values in  $\mathcal{X}$ . Formally,  $Z$  is an IV if, for any  $(x, z) \in \mathcal{X} \times \mathcal{Z}$ ,

$$\Pr(D_R = 1 \mid X = x, Z = z) \neq \Pr(D_R = 1 \mid X = x)$$

but

$$\Pr(D_Y = 1 \mid X = x, Z = z) = \Pr(D_Y = 1 \mid X = x).$$

Manski (1994, 2003) shows that if  $Z$  is an IV, then upper and lower bounds on the conditional poverty rate  $\Pr(D_Y = 1 \mid X = x)$  are

$$\begin{aligned} \text{UB}_{IV}(x) = \inf_z \{ & \Pr(D_Y = 1 \mid X = x, Z = z, D_R = 1) \Pr(D_R = 1 \mid X = x, Z = z) + \\ & + \Pr(D_R = 0 \mid X = x, Z = z) \}, \end{aligned}$$

$$\text{LB}_{IV}(x) = \sup_z \{ \Pr(D_Y = 1 \mid X = x, Z = z, D_R = 1) \Pr(D_R = 1 \mid X = x, Z = z) \}.$$

Further, these bounds are sharp. Although it is generally difficult to find valid instrumental variables, we believe that a convincing case can be made for data collection characteristics (characteristics of the interviewer, interview mode, length and design of the questionnaire, etc.), because they help predict nonresponse (see for examples Lepkowski and Couper 2002, Schr  pler 2004, and Nicoletti and Peracchi 2005) but lack predictive power for household income or poverty status.

Since IV restrictions are often controversial, another possibility is to impose the weaker monotone instrumental variable (MIV) restrictions. A random variable  $Z$  is a MIV if it shifts monotonically the poverty rate, possibly after conditioning on a set  $X$  of observable covariates. Formally,  $Z$  is a MIV if, for any  $x \in \mathcal{X}$ ,

$$\Pr(D_Y = 1 \mid X = x, Z = z) \geq \Pr(D_Y = 1 \mid X = x, Z = z') \quad (2)$$

whenever  $z \geq z'$  (or  $z \leq z'$ ). It is often easier to find a variable which is monotonically related to the outcome of interest than to find a proper IV. Manski and Pepper (2000) show that if  $Z$  is a MIV, then sharp bounds on the conditional poverty rate  $\Pr(D_Y = 1 \mid X = x, Z = z)$  are

$$\begin{aligned} \text{UB}_{MIV}(x, z) &= \inf_{z' \geq z} \{ \Pr(D_Y = 1 \mid X = x, Z = z', D_R = 1) \Pr(D_R = 1 \mid X = x, Z = z') + \\ &\quad + \Pr(D_R = 0 \mid X = x, Z = z') \}, \\ \text{LB}_{MIV}(x, z) &= \sup_{z' \leq z} \{ \Pr(D_Y = 1 \mid X = x, Z = z', D_R = 1) \Pr(D_R = 1 \mid X = x, Z = z') \}. \end{aligned}$$

Bounds on the population poverty rate  $\Pr(D_Y = 1)$  are simply obtained by averaging the conditional bounds  $\text{LB}_{IV}(x)$  and  $\text{UB}_{IV}(x)$  with respect to the distribution of  $X$ , or the conditional bounds  $\text{LB}_{MIV}(x, z)$  and  $\text{UB}_{MIV}(x, z)$  with respect to the joint distribution of  $(X, Z)$ .

As a third possibility, we suggest exploiting another source of information, namely the fact that nonrespondents may provide partial information on their income. In the ECHP, and many other surveys where household income is obtained by adding-up a number of different income components across household members, nonresponse to household income 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. This information on partially reported income provides a simple but effective way of shrinking the worst-case bounds, or the bounds obtained by imposing IV or MIV restrictions.

For example, 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

$$\begin{aligned} \Pr(D_Y = 1 | D_R = 0) &= \Pr(D_Y = 1 | D_{Y^*} = 1, D_R = 0) \Pr(D_{Y^*} = 1 | D_R = 0) + \\ &+ \Pr(D_Y = 1 | D_{Y^*} = 0, D_R = 0) \Pr(D_{Y^*} = 0 | D_R = 0), \end{aligned} \quad (3)$$

where  $D_{Y^*}$  equals one if  $Y^* \leq \gamma$  and equals zero otherwise. In the absence of measurement error,  $\Pr(D_Y = 1 | D_{Y^*} = 0, D_R = 0) = 0$  because partially reported income  $Y^*$  cannot exceed true income  $Y$ . Since the probability  $\Pr(D_Y = 1 | D_{Y^*} = 1, D_R = 0)$  must necessarily lie between zero and one, we obtain the following upper and lower bounds on the population poverty rate

$$\begin{aligned} \text{UB}^* &= \Pr(D_Y = 1 | D_R = 1) \Pr(D_R = 1) + \Pr(D_{Y^*} = 1 | D_R = 0) \Pr(D_R = 0), \\ \text{LB}^* &= \text{LB} = \Pr(D_Y = 1 | D_R = 1) \Pr(D_R = 1). \end{aligned}$$

Thus, the information on partially reported income provides a smaller upper bound 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  $\Pr(D_{Y^*} = 1 | D_R = 0) \Pr(D_R = 0)$ .

Our use of partially reported income to narrow the “worst-case” bounds is similar in spirit to the use of income bracket information by Vasquez-Alvarez et al. (1999, 2001) to bound income quantiles. They consider a sample survey where people who fail to provide their income are then asked to report whether their income exceeds a given threshold. We instead know that the income of nonrespondents is at least equal to partially reported income  $Y^*$ , which is not a fixed threshold but varies across individuals and can take any value between zero and  $Y$ .

### 3.2 Partial Identification in the Presence of Measurement Error

Measurement error in the poverty status occurs when either total household income or the household equivalent scale are measured with error. When the poverty line is also estimated, it may itself be affected by sampling noise or systematic bias.

If  $W$  denotes the observed (error-ridden) equivalized net income of a household and  $\hat{\gamma}$  denotes the estimated poverty line, then the observed poverty indicator  $D_W$  is equal to one if  $W \leq \hat{\gamma}$  and is equal to zero otherwise, and the observed poverty rate is  $\Pr(D_W = 1) = \Pr(W \leq \hat{\gamma})$ . When  $D_Y \neq D_W$ , poverty status is measured with error. Since  $D_Y$  and  $D_W$  are categorical indicators, the measurement error problem becomes a problem of misclassification that may arise either because  $Y \neq W$  due to measurement error in total household income or in the equivalence scale, or because

$\hat{\gamma} \neq \gamma$  due to sampling noise or systematic bias in the estimated poverty line. Ignoring the problem may lead to biased estimates of the population poverty rate  $\Pr(D_Y = 1)$ . An alternative approach, introduced by Horowitz and Manski (1995) and adopted by Chavez-Martin del Campo (2004), Pudney and Francavilla (2006), and Molinari (2003, 2008), is to partially identify  $\Pr(D_Y = 1)$  using the sample information along with weak assumptions about the measurement error process.

Horowitz and Manski (1995) model the observed outcome as a mixture of the true outcome and an unknown contaminating variable (the corrupted sampling model), and provide a general framework for partially identifying population parameters of interest by imposing a non-trivial upper bound on the probability of observing the contaminating variable. For a binary poverty indicator, their mixture model takes the form

$$D_W = D_Y(1 - D_*) + D_V D_*, \quad (4)$$

where  $D_*$  is equal to zero when we observe the true poverty indicator  $D_Y$  and is equal to one when we observe the contaminating binary indicator  $D_V$ .

Chavez-Martin del Campo (2004) specializes the results of Horowitz and Manski (1995) to poverty measures. By considering a mixture model for household income and by assuming a non-trivial upper bound on the measurement error probability, he shows how to bound poverty measures that are additively separable, a class which includes the headcount ratio.

Pudney and Francavilla (2006) also consider a mixture model for household income to investigate the effect of measurement error on estimation of poverty rates. Assuming that there are non-trivial levels of wellbeing at which people can be classified without error as poor or non-poor, that the contaminating variable does not depend on the level of wellbeing, and that the measurement error depends neither on the level of wellbeing nor on true or contaminated income (after conditioning on a set of variables), they show that one can exactly identify the poverty rate. They also show how to partially identify the poverty rate when some of these assumptions are relaxed.

An alternative approach, pioneered by Molinari (2003, 2008), is to directly bound the poverty rate by exploiting the identity

$$\Pr(D_W = 1) = \Pr(D_W = 1 \mid D_Y = 1) \Pr(D_Y = 1) + \Pr(D_W = 1 \mid D_Y = 0) \Pr(D_Y = 0). \quad (5)$$

This is just an implication of the law of total probability and places no restrictions on the relation

between the error-ridden indicator  $D_W$  and the error-free indicator  $D_Y$ . When coupled with assumptions about its elements, however, it generates a statistical model which Molinari (2008) calls a direct misclassification model. The main advantage of this approach is that it takes 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.

Molinari’s base-case assumptions are non-trivial upper bounds on either the overall misclassification probability  $\Pr(D_W \neq D_Y)$  or the direct misclassification probabilities  $\Pr(D_W = i \mid D_Y = j)$ , for  $i \neq j$ .

**Assumption B**  $\Pr(D_W \neq D_Y) \leq \lambda < 1$ .

**Assumption D**  $\Pr(D_W = i \mid D_Y = j) \leq \lambda < 1$ , for  $i \neq j$ .

Notice that Assumption D is stronger than Assumption B. In some cases, for example when validation studies are available, one may be able to directly estimate the upper bound  $\lambda$  in these two assumptions. Even when this is not possible, it may still be of interest to determine how inference about the population poverty rate changes with changes in the assumed bounds.

Proposition 3 in Molinari (2008) presents the bounds on the population poverty rate implied by the two assumptions. Assumption B gives

$$\text{UB}_B = \min\{\Pr(D_W = 1) + \lambda, 1\},$$

$$\text{LB}_B = \max\{\Pr(D_W = 1) - \lambda, 0\}.$$

These are the same bounds obtained by Horowitz and Manski (1995) under the assumption of an upper bound  $\lambda$  on  $\Pr(D_* = 1)$  in the mixture model (4). Assumption D gives instead

$$\begin{aligned} \text{UB}_D &= \min\left\{\frac{\Pr(D_W = 1)}{1 - \lambda}, 1\right\}, \\ \text{LB}_D &= \max\left\{\frac{\Pr(D_W = 1) - \lambda}{1 - \lambda}, 0\right\}. \end{aligned}$$

These are the same bounds obtained by Horowitz and Manski (1995) when replacing the mixture model (4) by a contaminated sampling model, namely one where  $D_Y$  and  $D_*$  are independent.

Molinari (2008) also shows how to identify narrower bounds by imposing additional restrictions on the direct misclassification probabilities. One such restriction is that the direct misclassification probabilities are constant, which together with Assumption D implies the following:

**Assumption CD**  $\Pr(D_W = 1 | D_Y = 0) = \Pr(D_W = 0 | D_Y = 1) \leq \lambda < 1$ .

Another restriction is monotonicity in correct reporting, that is,  $\Pr(D_W = j | D_Y = j) \geq \Pr(D_W = j + 1 | D_Y = j + 1)$ , which together with Assumption D implies the following:

**Assumption MD**  $\Pr(D_W = 1 | D_Y = 0) \leq \Pr(D_W = 0 | D_Y = 1) \leq \lambda < 1$ .

Assumption MD states that it is more likely for poor people to report an income above the poverty line than for rich people to report an income below the poverty line. This may possibly be the case when poverty (low income) is perceived by survey respondents as a stigma. The assumption that people underreport social undesirable characteristics is often made by survey methodologists and cognitive psychologists (see for example DeMaio 1984, Groves 1989, and Tourangeau et al. 2004). Assumption MD is also supported by several validations studies which find that measurement error in income is negatively correlated with true income (see for example Bound and Krueger 1991, Rodgers et al. 1993, Bound et al. 1994).

Although our approach is very similar in spirit to Molinari's direct misclassification approach, our starting point is neither the mixture model (4) nor the direct misclassification model (5). Instead, we consider the following relationship

$$\Pr(D_Y = 1) = \Pr(D_Y = 1 | D_W = 1) \Pr(D_W = 1) + \Pr(D_Y = 1 | D_W = 0) \Pr(D_W = 0). \quad (6)$$

Again, this is simply an implication of the law of total probability and imposes no restriction on the relation between the error-free and the error-ridden indicator of poverty. However, placing assumptions on its elements  $\Pr(D_Y = i | D_W = j)$  gives a new statistical model which we call an indirect misclassification model.

Given (6), an assumption that partially identifies the population poverty rate is the following:

**Assumption I**  $\Pr(D_Y = i | D_W = j) \leq \lambda < 1$ , for  $i \neq j$ .

While Assumption D restricts the conditional distribution of  $D_W$  given  $D_Y$  by placing an upper bound on the direct misclassification probabilities  $\Pr(D_W = i | D_Y = j)$ , for  $i \neq j$ , Assumption I restricts the conditional distribution of  $D_Y$  given  $D_W$  by placing an upper bound on the indirect misclassification probabilities  $\Pr(D_Y = i | D_W = j)$ , for  $i \neq j$ . It is easy to verify that,

while Assumption I implies Assumption B, there is no simple relation between Assumption I and Assumption D.

For expositional convenience, and without loss of generality, we use the same symbol  $\lambda$  for the upper bounds in Assumptions B, D and I, and the rest of our theoretical presentation. On the contrary, in our empirical application we allow  $\lambda$  to vary depending on the assumption considered.

The next proposition presents the bounds on the population poverty rate implied by Assumption I. To save space, all proofs are omitted but can be downloaded as Supplemental Materials from the *JBES* web site.

**Proposition 1** *If Assumption I holds, then*

$$UB_I = (1 - \lambda) \Pr(D_W = 1) + \lambda,$$

$$LB_I = (1 - \lambda) \Pr(D_W = 1).$$

*Further, these bounds are sharp.*

Figure 1 plots the upper and lower bounds implied by Assumptions B, D and I against  $\lambda$  for different values of  $\Pr(D_W = 1)$ . If  $\lambda = 0$ , then  $\Pr(D_Y = 1)$  is point-identified and coincides with  $\Pr(D_W = 1)$ . When  $\lambda > 0$ , the identification region implied by Assumption B contains those implied by Assumptions D and I. This is not surprising since Assumption B is weaker than Assumptions D and I.

Assumptions D and I are different, and there are no theoretical reasons to prefer one to the other. Their validity can be supported only by validation studies, while their usefulness in narrowing the bounds depends on the values of  $\lambda$  and  $\Pr(D_W = 1)$ . One important difference between the bounds based on Assumption D and those based on Assumption I is that, unlike the former, the latter are always informative (that is, they are different from zero and one whenever  $0 < \lambda < 1$ ) and change smoothly with  $\lambda$ . As for the width of the implied bounds,  $LB_D$  is always lower or equal to  $LB_I$  if  $\lambda > 1 - \Pr(D_W = 0)/\Pr(D_W = 1)$ , while  $UB_D$  is lower or equal to  $UB_I$  if  $\lambda < \Pr(D_W = 0)$  and  $\lambda > 1 - \Pr(D_W = 1)/\Pr(D_W = 0)$ . Moreover, if  $\lambda(1 - \lambda) \leq \Pr(D_W = 1) \leq 1 - \lambda(1 - \lambda)$ , then the interval identified by Assumption I is narrower than the one identified imposing Assumption D. On the contrary, if  $\Pr(D_W = 1)$  lies outside the interval  $[\lambda(1 - \lambda), 1 - \lambda(1 - \lambda)]$ , then Assumption D implies a narrower interval.



We also consider two additional assumptions, which represent the analogues of Assumptions CD and MD in Molinari (2008). The first is the assumption that the probability of indirect misclassification is constant:

**Assumption CI**  $\Pr(D_Y = 0 \mid D_W = 1) = \Pr(D_Y = 1 \mid D_W = 0) \leq \lambda < 1$ .

The second is the assumption that the probability of indirect misclassification is monotonic:

**Assumption MI**  $\Pr(D_Y = 0 \mid D_W = 1) \leq \Pr(D_Y = 1 \mid D_W = 0) \leq \lambda < 1$ .

The next result gives the identification intervals for the population poverty rate under these two assumptions.

**Proposition 2**

(i) *If Assumption CI holds, then*

$$\begin{aligned} \text{UB}_{CI} &= \begin{cases} (1 - 2\lambda) \Pr(D_W = 1) + \lambda, & \text{if } \Pr(D_W = 1) \leq 1/2, \\ \Pr(D_W = 1), & \text{otherwise,} \end{cases} \\ \text{LB}_{CI} &= \begin{cases} \Pr(D_W = 1), & \text{if } \Pr(D_W = 1) \leq 1/2, \\ (1 - 2\lambda) \Pr(D_W = 1) + \lambda, & \text{otherwise.} \end{cases} \end{aligned}$$

(ii) *If Assumption MI holds, then*

$$\text{UB}_{MI} = (1 - \lambda) \Pr(D_W = 1) + \lambda,$$

$$\text{LB}_{MI} = \text{LB}_{CI}.$$

Tables showing the identification intervals and their width under our Assumptions CI and MI, and the analogue Assumptions CD and MD in Molinari (2008), can be downloaded as Supplemental Materials from the *JBES* web site.

Following Manski and Pepper (2000) and Manski (2003), we may further narrow the bounds by imposing IV and MIV restrictions. Adopting the notation in Section 3.1, let  $Z$  be the IV or the MIV, let  $X$  be a set of covariates, and replace Assumptions B, D and I by the stronger assumptions:

**Assumption B\***  $\Pr(D_W \neq D_Y \mid X = x, Z = z) \leq \lambda < 1$  for any  $(x, z) \in (\mathcal{X} \times \mathcal{Z})$ .

**Assumption D\***  $\Pr(D_W = i \mid D_Y = j, X = x, Z = z) \leq \lambda < 1$  for  $i \neq j$  and any  $(x, z) \in (\mathcal{X} \times \mathcal{Z})$ .

**Assumption I\***  $\Pr(D_Y = i | D_W = j, X = x, Z = z) \leq \lambda < 1$  for  $i \neq j$  and any  $(x, z) \in (\mathcal{X} \times \mathcal{Z})$ .

Since these assumptions are stronger than Assumptions B, D, and I, in our application we choose higher values of  $\lambda$  when considering IV and MIV restrictions. Except for this, the basic idea is very simple. We first use these restrictions to bound the conditional poverty rate  $\Pr(D_Y = 1 | X = x, Z = z)$ , and then we obtain bounds on the population poverty rate  $\Pr(D_Y = 1)$  by averaging the conditional bounds with respect to the joint distribution of  $(X, Z)$ .

### 3.3 Partial Identification in the Presence of Missing Data and Measurement Error

In the presence of both missing data and measurement error, identification of the poverty rate becomes more problematic. 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 only observe a contaminated poverty indicator  $D_W$  instead of the unobserved indicator  $D_Y$ , while for nonresponding people we observe neither  $D_W$  nor  $D_Y$ .

The partial identification approaches discussed in Section 3.2 can be directly applied to find upper and lower bounds for  $\Pr(D_Y = 1 | D_R = 1)$ , the poverty rate for the respondents. All we need is an upper bound on either the misclassification probability, the direct misclassification probabilities, or the indirect misclassification probabilities, after conditioning on the event  $D_R = 1$ . For example, let BR denote Assumption B modified by conditioning on the event  $D_R = 1$ , and let  $LB_R$  and  $UB_R$  denote the implied upper and lower bounds on  $\Pr(D_Y = 1 | D_R = 1)$ , the poverty rate for the respondents. These are the same bounds obtained in Section 3.2, except that we now condition on the event  $D_R = 1$ . The resulting bounds on the unconditional poverty rate  $\Pr(D_Y = 1)$  are

$$UB_{BR} = UB_R \Pr(D_R = 1) + \Pr(D_R = 0),$$

$$LB_{BR} = LB_R \Pr(D_R = 1).$$

The same argument may be repeated for Assumptions D, I, CD, MD, CI and MI modified by conditioning on the event  $D_R = 1$ . In what follows, we denote these modified assumptions as DR, IR, CDR, MDR, CIR and MIR respectively.

When nonrespondents provide partial information on their income, these bounds can be narrowed further. If  $W^*$  is error-ridden partially-reported income and  $\hat{\gamma}$  is the estimated poverty line, then equation (3) must be modified as follows

$$\begin{aligned}\Pr(D_Y = 1 \mid D_R = 0) &= \Pr(D_Y = 1 \mid D_{W^*} = 1, D_R = 0) \Pr(D_{W^*} = 1 \mid D_R = 0) \\ &\quad + \Pr(D_Y = 1 \mid D_{W^*} = 0, D_R = 0) \Pr(D_{W^*} = 0 \mid D_R = 0).\end{aligned}$$

In the absence of measurement error, one can safely assume that  $\Pr(D_Y = 1 \mid D_{W^*} = 0, D_R = 0) = 0$ . In the presence of measurement error, this assumption is still quite reasonable because a household with partially reported income above the poverty line is unlikely to be poor. Under this assumption, it is enough to replace the term  $\Pr(D_R = 0)$  in  $UB_{BR}$ ,  $UB_{DR}$  and  $UB_{IR}$  by  $\Pr(D_R = 0) \Pr(D_{W^*} = 1 \mid D_R = 0)$ , leaving the lower bounds  $LB_{BR}$ ,  $LB_{DR}$  and  $LB_{IR}$  unchanged. In this case, the information on reported income causes the various identification regions to shrink by an amount equal to  $\Pr(D_R = 0) [1 - \Pr(D_{W^*} = 1 \mid D_R = 0)]$ .

Computation of the bounds using IV and MIV is straightforward after conditioning Assumptions  $B^*$ ,  $D^*$  and  $I^*$  on the event  $D_R = 1$ .

## 4 EMPIRICAL RESULTS

We now present the estimated bounds for the population poverty rates based on the results in Section 3. These bounds are computed considering both measurement error and missing data problems. In Section 4.1, we derive bounds by first imposing an upper bound on the misclassification probabilities and by then imposing the additional assumption of monotonicity in correct reporting. In Section 4.2 we study how the identification intervals for the poverty rates change when we choose different upper bounds on the misclassification probabilities. Finally, in Section 4.3 we impose additional IV and MIV assumptions.

### 4.1 Bound Estimates

This section presents, separately by country, the estimated bounds for the population poverty rates. These bounds are functions of probabilities which are estimated nonparametrically by simple weighted empirical frequencies using the survey weights provided by the ECHP. Since the bounds

are estimated, we also take their sampling variability into account. This is done by constructing 90%-level bootstrap confidence intervals based on the percentile method and 1,000 bootstrap replications. These confidence intervals cover the entire identification region with 90% probability. Unlike standard asymptotic confidence intervals, they are generally not symmetric. The bootstrap samples are obtained by sampling with replacement households, not individuals. Further, for each bootstrap sample, the cross-sectional weights are rescaled to have unit mean (Biewen 2002).

The choice of the upper bounds for the misclassification probabilities are based on the validation study of Epland and Kirkeberg (2002), who compare true and reported income by matching administrative data with the 1996 Norwegian Survey of Living Conditions. Using their Table 1, and setting the poverty line at 100,000 Norwegian crowns (which roughly corresponds to 60% of median equivalized household income), we find that the estimated probability that true and reported poverty status differ (the misclassification probability) is about 6.5 percent. The estimated direct, indirect and overall misclassification probabilities, and their standard error, are shown in Table 2.

Assumption MD of monotonicity in correct reporting is confirmed by the results in Table 2, whereas the assumption MI is not. Results hardly change when increasing or decreasing the poverty line by 50 percent. Thus, in our empirical application, we consider the following assumptions:

**Assumption BR**  $\Pr(D_W \neq D_Y | D_R = 1) \leq \lambda_{BR}$ .

**Assumption MDR**  $\Pr(D_W = 1 | D_Y = 0, D_R = 1) \leq \Pr(D_W = 0 | D_Y = 1, D_R = 1) \leq \lambda_{MDR}$ .

**Assumption IR**  $\Pr(D_Y = i | D_W = j, D_R = 1) \leq \lambda_{IR}$ , for  $i \neq j$ .

The bounds on the misclassification probabilities are set to the estimated values in Table 2 plus twice their standard error, i.e.  $\lambda_{BR} = 0.073$ ,  $\lambda_{MDR} = 0.113$  and  $\lambda_{IR} = 0.140$ . In Section 4.2 we also conduct a sensitivity analysis to study how results change when we vary the upper bounds.

Table 3 reports the estimated upper and lower bounds on the population poverty rate and the corresponding upper and lower limits of their bootstrap confidence interval under the three assumptions. We denote the three identification intervals as  $[\text{LB}_{BR}^*, \text{UB}_{BR}^*]$ ,  $[\text{LB}_{DR}^*, \text{UB}_{DR}^*]$ , and  $[\text{LB}_{IR}^*, \text{UB}_{IR}^*]$  (the superscript \* indicates that partially reported income is used to compute the

bounds). Upper bounds tend to be lower under Assumption MDR than under Assumptions IR and BR, whereas lower bounds are higher under Assumption IR than under Assumption MDR and BR. Assumption MDR produces the narrowest bounds, whose length goes from 0.136 for Denmark to 0.164 for the UK. This is unsurprising since Assumption MDR combines Assumption DR and the monotonicity assumption.

If all three Assumptions BR, IR and MDR hold at the same time, then we can compute narrower bounds. The resulting identification interval for the population poverty rate is denoted by  $[LB_J^*, UB_J^*]$ , where  $LB_J^*$  is the maximum between  $LB_{BR}^*$ ,  $LB_{MDR}^*$  and  $LB_{IR}^*$ , while  $UB_J^*$  is the minimum between  $UB_{BR}^*$ ,  $UB_{MDR}^*$  and  $UB_{IR}^*$ . Estimates of this new set of bounds are presented in Table 4. The range of plausible values is reduced considerably, as the width now varies between 0.055 (0.089 in terms of bootstrap confidence intervals) and 0.101 (0.186). Although the estimated identification regions overlap partially for several countries some clear results emerge. In Denmark the estimated upper bound on the poverty rate is lower than the lower bounds estimated for Greece, Ireland, Italy, Portugal, Spain and the UK. Similarly, the Netherlands has an estimated upper bound which is lower than the ones estimated for Greece, Ireland, Italy and Portugal; the estimated upper bound for Germany is lower than for Ireland and Portugal; and the one for Belgium is lower than for Portugal. Based on these results we can reject the hypotheses that poverty rates in Belgium, Denmark, Germany and the Netherlands are higher than in the remaining countries.

Furthermore, by ranking countries in terms of their upper bound on the poverty rate, 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, Spain, Ireland and the UK make up an intermediate group. Ireland moves from the intermediate group to the high-poverty group if we rank the countries using the lower bound. Interestingly, this is in line with the country ranking obtained using the point estimates of poverty rates in Section 2, with Ireland being positioned between the high-poverty group and the intermediate one.

Table 4 also presents a decomposition of the width  $\Delta = UB_J^* - LB_J^*$  of the identification region into two additive components. The first component,  $\Delta_1 = \Pr(D_R = 0) \Pr(D_{W^*} = 1 | D_R = 0)$ , is caused by the presence of missing data. The second component,  $\Delta_2 = UB_J - LB_J - \Delta_1$ , is instead caused by measurement errors affecting the observed poverty indicator. For all countries, at most

33.2% of the interval width is determined by the presence of measurement errors problems. This suggests that the lack of identification is mainly due to missing data problems, at least for the values of  $\lambda$  that have been chosen for this application.

## 4.2 Sensitivity Analysis

Even if based on validation studies, the choice of upper bounds on the misclassification probabilities is to some extent arbitrary. Thus, we also carry out a sensitivity analysis by looking at how results change when we allow these upper bounds to vary. We compute for each country the width of  $[LB_J^*, UB_J^*]$  (the intersection between  $[LB_{BR}^*, UB_{BR}^*]$ ,  $[LB_{IR}^*, UB_{IR}^*]$  and  $[LB_{MDR}^*, UB_{MDR}^*]$ ) for different values of  $\lambda_{BR}$ ,  $\lambda_{MDR}$  and  $\lambda_{IR}$ . More precisely, we allow the upper bound of the indirect misclassification probability,  $\lambda_{IR}$ , to change from 0.01 to 0.99 and the upper bounds  $\lambda_{MDR}$  and  $\lambda_{BR}$  to vary proportionally with  $\lambda_{IR}$ . We keep the ratio between  $\lambda_{MDR}$  ( $\lambda_{BR}$ ) and  $\lambda_{IR}$  equal to the ratio between 0.113 (0.073) and 0.140, which are the values used in the previous section.

In presenting the results, we focus on the width of the intervals defined by the estimated bounds because it is a measure of how serious the identification problem is. A zero width corresponds to point identification of the true poverty rate, while a width that is positive but less than one corresponds to partial identification.

Table 5 reports the minimum and the maximum widths over all countries of the estimated interval  $[LB_J^*, UB_J^*]$  for different values of the  $\lambda$ 's. Both the minimum and the maximum widths increase with  $\lambda$ . The widths are always smaller than .251 for values of  $\lambda_{IR}$  less than or equal to .95. Of the three assumptions, MDR produces the lowest upper bound when  $\lambda_{MDR}$  is less than  $1 - \Pr(D_W = 1, D_R = 1)$ , while BR produces the highest lower bound. It is only when  $\lambda_{MDR} > 1 - \Pr(D_W = 1, D_R = 1)$  that Assumption IR produces a lower upper bound than Assumption MDR, and this happens only for Italy and Portugal when  $\lambda_{MDR}$  is fixed at its highest value.

These results may be useful to survey methodologists interested in improving the quality of a survey by adopting techniques aimed at reducing nonresponse rates or measurement errors. For example, from these results, it seems that the missing data problem is the main cause of lack of identification. When  $\lambda_{IR} \leq 0.2$ , the missing data problem is always the main explanation for the lack of identification. When  $\lambda_{IR} = 0.5$ , there are still countries where the missing data problem is

the main explanation for the lack of identification. Even when  $\lambda_{IR} = 0.99$ , we still find that the missing data problem explains between 10% and 38.7% of the interval length. We can reject the assumption that the missing data problem is the main explanation for the lack of identification only when we assume that  $\lambda_{IR} = 0.500$ ,  $\lambda_{MDR} = 0.404$  and  $\lambda_{BR} = 0.261$ . Notice that these values are more than three times higher than the corresponding misclassification probabilities found in the validation studies of Epland and Kirkeberg (2002) (See Table 2). For this reason, we conclude that measurement error is of secondary importance relative to missing data in our empirical application.

### 4.3 Restricting the Bounds Using IV and MIV Assumptions

When IV and MIV restrictions are introduced, estimation of the bounds is complicated by issues of finite-sample bias, due to the small size of the cells over which we impose these assumptions. As shown by Kreider and Pepper (2007), sample estimates based on infima and suprema will be systematically biased and the estimated bounds will be too narrow, so we correct the estimates and the confidence interval using the bootstrap bias correction that they propose. The Monte Carlo experiments conducted by Manski and Pepper (2009) to study the small sample properties of this correction show that the bias reduces considerably and becomes negligible.

In our application, we explored various IV candidates—in particular variables related to the data collection process—by testing their statistical significance in a probit model for the response probability. In the end, our best choice is the total number of successful interviews in the previous waves. We use this variable as an IV after controlling for household size, the number of workers, the number of children, and the education level of the reference person.

As MIV's, we consider the size of the household and the number of its working members. We use them as alternative MIV's after controlling for the number of children, the reference person's education, and, in addition, either the number of working household members (for the former MIV) or the household size (for the latter). Thus, we replace Assumptions BR, MDR and IR in Section 4.1 by analogues based on the assumed IV or MIV and the additional covariates. For example Assumption BR, is replaced by the more restrictive assumption that  $\Pr(D_W \neq D_Y \mid D_R = 1, X = x, Z = z) \leq \lambda$  for any  $(x, y) \in \mathcal{X} \times \mathcal{Z}$ , where  $Z$  is the IV or the MIV and  $X$  contains the control variables. We proceed in the same way with Assumptions MDR and IR.

Because misclassification probabilities may depend on  $X$  and  $Z$ , one may in principle consider

different upper bounds for each  $x$  and  $z$  value. This approach is not feasible due to the lack of validation studies reporting misclassification probabilities by household size, number of children, etc. For this reason, we fix a common upper bound valid for any  $x$  and  $z$  value. This upper bound is equal to the largest misclassification probability estimated by Epland and Kirkeberg (2002) plus four times the standard error of this estimate. Although arbitrary, the choice of multiplying the standard errors by four is for caution.

Table 7 presents the estimated bounds, separately for our IV and MIV restrictions, under the assumption that BR, MDR and IR all hold when conditioning on the IV or MIV and the additional covariates. We exclude the Netherlands from the analysis because the quality of the data on education is doubtful. The narrowest bounds are those identified by the stronger IV restriction, and their widths vary from 0.02 (0.04 for the bootstrap confidence interval) to 0.07 (0.09). For all countries except Portugal, the intervals identified by these bounds are narrower than those obtained under the joint assumptions BR, MDR and IR.

By comparing the estimated bounds across countries, we draw the following conclusions: (1) Belgium, Denmark and Germany are the countries with the lowest poverty rates ( $UB_{IV}^*$  for these countries is lower than  $LB_{IV}^*$  for all other countries); (2) Italy and Portugal have higher poverty rates than Ireland, Spain and the UK ( $LB_{IV}$  for Italy is higher than  $UB_{IV}$  for Ireland and the UK, while  $UB_{MIV_2}$  for Spain and  $LB_{MIV_1}$  for Portugal are higher than  $UB_{MIV_1}$  for Ireland and  $UB_{IV}$  for the UK); (3) Greece has a higher poverty rate than Ireland ( $LB_{MIV_1}$  for Greece is higher than  $UB_{MIV_2}$  for Ireland). If we look at the confidence interval, then the identification regions become slightly larger and this weakens some of our conclusions. Nevertheless, our results suggest the presence of three groups of countries with different levels of poverty: low for Belgium, Denmark and Germany, medium for Ireland, Spain and the UK, and high for Greece, Italy and Spain.

## 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 the overall misclassification probability, the direct misclassification probability, or the indirect misclassification



probability. While assumptions on the existence of an upper bound on the misclassification probability and on the direct misclassification probability have already been used to partially identify probability distributions (see for example Horowitz and Manski 1995 and Molinari 2008), we are the first to use assumptions on the indirect misclassification probability. Furthermore, we show how to extend the partial identification approach to the case where measurement error and missing data problems coexist, and how to use assumptions on misclassification probabilities together with instrumental variables and monotone instrumental variables assumptions.

By applying this extended partial identification approach, we estimate upper and lower bounds for the poverty rates in 10 European countries. Our main results can be summarized as follows. First, the use of assumptions on misclassification probabilities jointly with IV and MIV restrictions are very useful in partial identification of poverty rates. In our empirical application, these assumptions allow us to identify bounds which are narrow enough to be informative about the ranking of countries by level of poverty.

Second, in the presence of both measurement errors and missing data, partial identification provides information on which of the two problems survey methodologists and applied social scientists should be more concerned with. This is possible by decomposing the identification intervals into the part due to the missing data and that due to the measurement error. To reject the assumption that the missing data problem is the main explanation for the lack of identification, we have to increase the upper bounds on the misclassification probabilities to values which are much larger than those observed in validation studies. We conclude that missing data should be the major concern when estimating poverty rates using surveys similar to the ECHP.

## ACKNOWLEDGEMENTS

We thank Christopher Bollinger, Chuck Manski, Francesca Molinari, Tom Wansbeek, a Joint Editor, an Associate Editor, and two anonymous referees for very helpful suggestions. Cheti Nicoletti's research was supported by the Economic and Social Research Council through their grant to the Research Centre on Micro-Social Change at the ISER. Part of this paper is based on research carried out during Francesca Foliano's visit to the European Centre for Analysis in the Social Sciences (ECASS) at the ISER, supported by the Access to Research Infrastructure action under the EU

Improving Human Potential Programme.

## SUPPLEMENTAL MATERIALS

**Proofs:** Proofs of Theorems 1 and 2.

**Supplemental Table 1:** Upper and lower bounds under Assumptions CD, MD, CI and MI for  $\lambda \leq 1/2$  and different values of  $p = \Pr(D_W = 1)$ .

**Supplemental Table 2:** Upper and lower bounds under Assumptions CD, MD, CI and MI for  $\lambda \geq 1/2$  and different values of  $p = \Pr(D_W = 1)$ .

All supplemental items are contained in a single pdf file.

## REFERENCES

- Biewen, M. (2002), “Bootstrap Inference for Inequality, Mobility and Poverty Measurement,” *Journal of Econometrics*, 108, 317–342.
- Bound, J., and Krueger, A. (1991), “The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?,” *Journal of Labor Economics*, 9, 1–24.
- Bound, J., Brown, C., Duncan G. J., and Rodgers, W. L. (1994), “Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data,” *Journal of Labor Economics*, 12, 345–368.
- Bound, J., Brown, C., and Mathiowetz, N. (2001), “Measurement Error in Survey Data,” in *Handbook of Econometrics*, Vol. 5, eds. J. J. Heckam and E. Leamer, Amsterdam: North Holland, pp. 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., and 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.
- Deaton, A. (1997): *The Analysis of Household Surveys. A Microeconometric Approach to Development Policy*, Baltimore: Johns Hopkins University Press.
- DeMaio, T. J. (1984), “Social Desirability and Survey Measurement: A Review,” in *Surveying Subjective Phenomena*, eds. C. F. Turner and E. Martin, New York: Russell Sage, pp. 257–282.
- Epland, J., and 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, August 25–28, 2002.
- Eurostat (2003), “Imputation of Income in the ECHP,” PAN 164/01.
- Groves, R. M. (1989): *Surveys Errors and Survey Costs*, New York: Wiley.
- Horowitz, J. L., and Manski, C. F. (1995), “Identification and Robustness with Contaminated and Corrupted Data,” *Econometrica*, 63, 281–302.
- Horowitz, J. L., and 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., and Pepper, J. V. (2007), Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors,” *Journal of the American Statistical Association*, 101, 432–441.
- Little, J. A., and Rubin D. B. (1987): *Statistical Analysis with Missing Data*, New York: Wiley.
- Manski, C. F. (1989), “Anatomy of the Selection Problem,” *Journal of Human Resources*, 24, 343–360.
- Manski, C. F. (1994), “The Selection Problem,” in *Advances in Econometrics, Sixth World Congress*, ed. C. Sims, Cambridge: Cambridge University Press, pp. 143–170.
- Manski, C. F. (2003): *Partial Identification of Probability Distributions*, New York: Springer.

- Manski, C. F., and Pepper, J. (2000), "Monotone Instrumental Variables with an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010.
- Manski, C. F., and Pepper, J. (2009), "More on Monotone Instrumental Variables," *Econometric Journal*, forthcoming.
- Molinari, F. (2003), "Contaminated, Corrupted and Missing Data," unpublished Ph.D. Dissertation, Northwestern University, Department of Economics.
- Molinari, F. (2008), "Partial Identification of Probability Distributions with Misclassified Data," *Journal of Econometrics*, 144, 81–117.
- Nicoletti, C. (2009), "Poverty Analysis with Item and Unit Nonresponses: Alternative Estimators Compared," *Empirical Economics*, forthcoming.
- Nicoletti, C., and Peracchi F. (2002), "A Cross-Country Comparison of Survey Participation in the ECHP," ISER Working Paper 2002-32, University of Essex.
- Nicoletti, C., and Peracchi, F. (2006), "The Effects of Income Imputation on Microanalyses: Evidence from the European Community Household Panel," *Journal of Royal Statistical Society A*, 169, 625–1271.
- Pudney, S., and 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.
- Rodgers, W., Brown, C., and Duncan G. J. (1993), "Errors in Survey Reports of Earnings, Hours Worked, and Hourly Wages," *Journal of the American Statistical Association*, 88, 1208–1218.
- Rubin, D. B. (1989): *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.
- Rubin, D. B. (1996), "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, 91, 473–520.
- Schräpler, J.-P. (2004), "Respondent Behavior in Panel Studies – A Case Study for Income-Nonresponse by Means of the German Socio-Economic Panel (SOEP)," *Sociological Methods & Research*, 33, 118–156.
- van Praag, B., Hagenaars, A., and van Eck, W. (1983), "The Influence of Classification and Observation Errors on the Measurement of Income Inequality," *Econometrica*, 51, 1093–1108.
- Tourangeau, R., Lance, J. R., and Rasinski, K. (2004): *The Psychology of Survey Response*, Cambridge: Cambridge University Press.
- Vasquez-Alvarez, V. R., Melenberg, B., and 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., and van Soest, A. (2001), "Nonparametric Bounds in the Presence of Item Nonresponse, 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: Estimated poverty rates and nonresponse rates by country in 2001 (standard errors in parentheses).

Country	No. obs.	Poverty rate with imputation	Poverty rate for respondents	Nonresponse rate
Belgium	5607	0.116 (0.005)	0.127 (0.006)	0.201 (0.005)
Denmark	4975	0.110 (0.009)	0.101 (0.008)	0.144 (0.005)
Germany	13489	0.111 (0.004)	0.109 (0.005)	0.157 (0.003)
Greece	11114	0.192 (0.004)	0.195 (0.004)	0.131 (0.003)
Ireland	5421	0.185 (0.008)	0.194 (0.008)	0.099 (0.004)
Italy	15317	0.195 (0.004)	0.211 (0.005)	0.190 (0.003)
Netherlands	10395	0.116 (0.004)	0.109 (0.004)	0.073 (0.003)
Portugal	12917	0.211 (0.007)	0.222 (0.007)	0.138 (0.003)
Spain	13689	0.172 (0.004)	0.173 (0.004)	0.123 (0.003)
UK	10681	0.160 (0.004)	0.165 (0.004)	0.102 (0.003)

Table 2: Misclassification probabilities in Epland and Kirkeberg (2002).

	Estimated value	S.E.
$\Pr(D_W \neq D_Y)$	0.065	0.004
$\Pr(D_W = 1 \mid D_Y = 0)$	0.052	0.005
$\Pr(D_W = 0 \mid D_Y = 1)$	0.094	0.009
$\Pr(D_Y = 1 \mid D_W = 0)$	0.041	0.004
$\Pr(D_Y = 0 \mid D_W = 1)$	0.119	0.010

Table 3: Estimated bounds by country. 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	$LB_{BR}^*$	$UB_{BR}^*$	$LB_{MDR}^*$	$UB_{MDR}^*$	$LB_{IR}^*$	$UB_{IR}^*$
Belgium	0.043	0.214	0.012	0.156	0.086	0.253
	0.027	0.248	0.000	0.188	0.072	0.289
Denmark	0.024	0.199	0.000	0.136	0.074	0.244
	0.003	0.263	0.000	0.198	0.054	0.311
Germany (SOEP)	0.031	0.212	0.000	0.150	0.080	0.256
	0.020	0.238	0.000	0.175	0.070	0.283
Greece	0.107	0.299	0.081	0.235	0.147	0.334
	0.094	0.324	0.067	0.259	0.135	0.360
Ireland	0.109	0.290	0.083	0.225	0.151	0.326
	0.084	0.339	0.055	0.272	0.127	0.378
Italy	0.113	0.298	0.091	0.238	0.149	0.329
	0.099	0.326	0.075	0.265	0.135	0.358
Netherlands	0.034	0.210	0.000	0.142	0.087	0.258
	0.023	0.235	0.000	0.166	0.077	0.284
Portugal	0.124	0.321	0.103	0.260	0.159	0.351
	0.099	0.384	0.076	0.321	0.136	0.417
Spain	0.085	0.277	0.057	0.215	0.126	0.313
	0.072	0.311	0.044	0.248	0.113	0.350
UK (BHPS)	0.083	0.283	0.053	0.217	0.128	0.322
	0.072	0.307	0.040	0.240	0.118	0.348

*Note:*  $BR$  ( $DR$  and  $IR$ ) stands for the assumption that the overall (the direct and the indirect) misclassification probability is lower than 0.073 (0.113 and 0.140). The superscript \* indicates that the bounds are computed using information on partial reported income.

Table 4: Estimates of  $UB_J$ ,  $LB_J$  and of the width  $\Delta = UB_J - LB_J$  by country. 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.  $\Delta_1$  is the part of the interval width due to missing data problems, while  $\Delta_2$  is that due to measurement error problems.

Country	$LB_J^*$	$UB_J^*$	Width= $\Delta$	$\Delta_1/\Delta$	%	$\Delta_2/\Delta$	%
Belgium	0.086	0.156	0.070	0.056	79.8	0.014	20.2
	0.072	0.188	0.117				
Denmark	0.074	0.136	0.062	0.050	80.6	0.012	19.4
	0.054	0.198	0.144				
Germany (SOEP)	0.080	0.150	0.071	0.058	81.7	0.013	18.3
	0.070	0.175	0.106				
Greece	0.147	0.235	0.088	0.064	72.7	0.024	27.3
	0.135	0.259	0.124				
Ireland	0.151	0.225	0.074	0.049	66.8	0.025	33.2
	0.127	0.272	0.145				
Italy	0.149	0.238	0.090	0.066	73.1	0.024	26.9
	0.135	0.265	0.130				
Netherlands	0.087	0.142	0.055	0.041	74.3	0.014	25.7
	0.077	0.166	0.089				
Portugal	0.159	0.260	0.101	0.075	74.4	0.026	25.6
	0.136	0.321	0.186				
Spain	0.126	0.215	0.089	0.068	76.8	0.021	23.2
	0.113	0.248	0.135				
UK (BHPS)	0.128	0.217	0.088	0.068	76.4	0.021	23.6
	0.118	0.240	0.122				

Table 5: Minimum and maximum width  $\Delta = \text{UB}_J - \text{LB}_J$  across countries for different values of  $\lambda_{IR}$  and  $\lambda_{MDR}$ .  $\Delta_1/\Delta$  is the part of the width due to missing data problems over the total width.

$\lambda_{BR}$	$\lambda_{MDR}$	$\lambda_{IR}$	min width	max width	min $\Delta_1/\Delta$	max $\Delta_1/\Delta$	mean $\Delta_1/\Delta$
0.005	0.008	0.010	0.042	0.077	0.966	0.984	0.977
0.026	0.040	0.050	0.046	0.084	0.849	0.926	0.896
0.052	0.081	0.100	0.051	0.094	0.738	0.862	0.813
0.104	0.161	0.200	0.061	0.112	0.585	0.757	0.686
0.156	0.242	0.300	0.071	0.131	0.484	0.675	0.594
0.261	0.404	0.500	0.092	0.168	0.360	0.555	0.469
0.365	0.565	0.700	0.110	0.205	0.287	0.471	0.388
0.417	0.646	0.800	0.119	0.223	0.260	0.438	0.357
0.469	0.726	0.900	0.128	0.242	0.238	0.410	0.331
0.495	0.767	0.950	0.132	0.251	0.229	0.397	0.319
0.516	0.799	0.990	0.135	0.688	0.100	0.387	0.274

Table 6: Means of the instrumental and monotone instrumental variables and the control variable

Country	IV	$\text{MIV}_1 \leq 2$	$\text{MIV}_1 = 3$	$\text{MIV}_2 = 0$	$\text{MIV}_2 = 1$	$x$
Belgium	0.767	0.351	0.188	0.227	0.244	0.305
Denmark	0.699	0.444	0.176	0.161	0.242	0.188
Germany	0.750	0.364	0.223	0.164	0.297	0.183
Greece	0.777	0.252	0.215	0.187	0.317	0.596
Ireland	0.871	0.21	0.151	0.157	0.272	0.538
Italy	0.672	0.216	0.261	0.148	0.332	0.626
Portugal	0.766	0.254	0.255	0.15	0.245	0.684
Spain	0.722	0.252	0.22	0.172	0.323	0.608
UK	0.762	0.379	0.206	0.217	0.284	0.405

*Note:* IV is the dummy variable for households that participated in the survey for at least 7 waves;  $\text{MIV}_1$  is the household size (the excluded category is a household of size greater than 3);  $\text{MIV}_2$  is the number of working household members (the excluded category is 2 or more); and  $x$  is the control for lower education.



Table 7: Estimated bounds by country. For each country, the estimates of the upper (lower) bounds are reported in the first row, the estimated finite-sample bias is reported in the second row while the corresponding upper (lower) limits of the corrected bootstrapped 90% confidence intervals are reported in the third row.

Country	LB <sub>IV</sub> *	UB <sub>IV</sub> *	LB <sub>MIV<sub>1</sub></sub> *	UB <sub>MIV<sub>1</sub></sub> *	LB <sub>MIV<sub>2</sub></sub> *	UB <sub>MIV<sub>2</sub></sub> *
Belgium	0.088	0.129	0.085	0.155	0.081	0.154
	0.007	-0.007	0.004	-0.008	0.001	-0.003
	0.078	0.141	0.075	0.168	0.069	0.171
Denmark	0.083	0.100	0.068	0.113	0.061	0.122
	0.003	-0.005	0.002	-0.011	0.001	-0.009
	0.069	0.111	0.060	0.126	0.052	0.132
Germany	0.119	0.139	0.087	0.154	0.083	0.159
	0.006	-0.006	0.002	-0.005	0.001	-0.003
	0.108	0.149	0.080	0.164	0.076	0.169
Greece	0.168	0.240	0.184	0.228	0.160	0.202
	0.008	-0.008	0.005	-0.003	0.002	-0.007
	0.160	0.252	0.166	0.258	0.151	0.233
Ireland	0.174	0.202	0.160	0.192	0.126	0.183
	0.007	-0.015	0.007	-0.003	0.001	-0.003
	0.140	0.218	0.130	0.217	0.111	0.206
Italy	0.204	0.232	0.162	0.261	0.160	0.254
	0.003	-0.002	0.003	-0.004	0.001	-0.013
	0.188	0.246	0.151	0.272	0.150	0.268
Portugal	0.176	0.257	0.204	0.255	0.168	0.262
	0.007	-0.008	0.003	-0.002	0.001	-0.008
	0.168	0.265	0.184	0.270	0.157	0.273
Spain	0.149	0.211	0.127	0.217	0.128	0.185
	0.004	-0.008	0.003	-0.005	0.001	-0.005
	0.133	0.221	0.119	0.231	0.117	0.203
UK	0.158	0.199	0.125	0.222	0.126	0.219
	0.006	-0.004	0.003	-0.004	0.000	0.000
	0.143	0.211	0.118	0.234	0.116	0.233

*Note:* The overall (the direct and the indirect) misclassification probabilities are assumed to be lower than 0.081 (0.130 and 0.159).

Figure 1: Bounds on the population poverty rate under Assumptions B, D and I as functions of  $\lambda$  for different values of  $\Pr(D_W = 1)$ .

