How much confidence can we have in EU-SILC? Complex sample designs and the standard error of the Europe 2020 poverty indicators

Tim Goedemé

Please cite the final version of this article as follows:

Goedemé, T. (2013). How much confidence can we have in EU-SILC? Complex sample designs and the standard error of the Europe 2020 poverty indicators. *Social Indicators Research*, 110(1), 89-110.

DOI: http://dx.doi.org/10.1007/s11205-011-9918-2.

More information on this topic can be found on my personal website:

https://timgoedeme.com/eu-silc-standard-errors/

1 Introduction

On 17 June 2010 the European Council agreed to reduce the number of Europeans at-risk-of-poverty or social exclusion by at least 20 million (European Council 2010). The target is defined by a composition of three different indicators: an indicator of financial poverty (the so-called at-risk-of-poverty rate), an indicator of material deprivation and an indicator of the number of households with a very low work intensity (details below). The data underlying these indicators are the European Union Statistics on Income and Living Conditions (EU-SILC), the principal data source for cross-national comparative research on income and living conditions in the European Union (EU). As EU-SILC is composed of samples in all EU Member States, sampling and non-sampling errors can seriously affect the accuracy of all estimates based on this survey – including the Europe 2020 poverty indicators. However, until now, Eurostat has refrained from consistently publishing standard errors and confidence intervals alongside the official poverty indicators (e.g. Eurostat 2010b; Wolff 2010). Unfortunately, this is not a feature unique to Eurostat publications. It seems to be rather common practice to ignore the publication of confidence intervals in the case of descriptive (poverty) statistics (e.g. OECD 2008; Kangas and Ritakallio 2007; de Vos and Zaidi 1998; Whelan and Maître 2007). Furthermore, in the case of analytical studies using a variety of (regression) methods, it is not always clear whether standard errors are calculated accurately.

Confidence intervals do not address all kinds of survey errors (e.g. Groves et al. 2009; Verma et al. 2010). Nevertheless, the estimation of confidence intervals can save money, time and effort. By showing whether differences between point estimates have a high probability of being due to random error, they indicate which differences are not worth further investigation. However, they can only serve this purpose if standard errors have been estimated accurately. In order to do so, among others it is necessary to take account of the sample design and weighting schemes. As is the case for many surveys (e.g. the World Values Survey (WVS), the Survey of Health, Ageing and Retirement in Europe (SHARE) and the Current Population Survey (CPS)), many countries covered by EU-SILC employ complex sample designs involving multiple stages of selection, stratification and clustering. However, as is also the case for other databases (such as the Luxembourg Income Study Database), full documentation of the sample design and accurate sample design variables in the EU-SILC dataset are lacking for a large number of participating countries. Therefore, in this article I explore the overall precision of EU-SILC and document the effect of various assumptions with regard to the sample design on the estimated standard errors. Focusing on the official Europe 2020 poverty reduction indicators, I show that researchers should at least take account of clustering within households when estimating standard errors. Although doing so leads in many cases to an acceptable approximation of

the standard error, it is found that by making optimal use of the sample design variables in the EU-SILC dataset, estimated standard errors are in most cases more accurate.

The article is structured as follows. First, I elaborate on the general principles of the computation of standard errors. Second, drawing on the reports published by Eurostat as well as personal correspondence with the national statistical institutes across Europe, I concisely discuss the sample design of EU-SILC and the available sample design variables. Third, I discuss the Europe 2020 poverty indicators and their associated standard errors. In the following section, I illustrate the effect on the standard error of making various simplifying assumptions with regard to the sample design for the Europe 2020 poverty indicators. These assumptions (e.g. simple random sampling) are often implicit in the standard errors as routinely produced by standard statistical software, and reported by poverty analysts. The resulting standard errors are compared to those obtained by using a dataset prepared by Eurostat which contains more complete information on the sample design. In the final section I discuss the wider application of the findings.

2 The estimation of standard errors: some principles

There are several approaches to the estimation of standard errors and the computation of confidence intervals (Heeringa et al. 2010; Wolter 2007). Standard errors and confidence intervals can be derived analytically, e.g through a technique called linearisation, and assuming a certain sampling distribution, usually Student's t-distribution or the normal distribution. A completely different approach is based on re-sampling from the original sample a high number of samples in order to empirically derive a sampling distribution (e.g. Jackknife repeated replication or the bootstrap). Subsequently, on the basis of this 'empirical sampling distribution' standard errors and confidence intervals are computed (cf. Mooney and Duval 1993 for an introduction; and Biewen 2002; Trede 2002; Van Kerm 2002; Davidson and Flachaire 2007; and del Mar Rueda and Muñoz 2011 for an application to poverty and inequality measures). There are various methods in between (see Efron and Tibshirani 1998) and each of these methods has its advantages and shortcomings. This section will not go into the details of the various approaches to variance estimation. Rather, it aims at summarising the general principles which always should be taken into account when computing standard errors. Whichever approach is used, in order to get the standard errors right one should replicate as closely as possible the entire procedure of drawing the sample and calculating the desired statistic. There are four main ingredients to this: sample design, weighting, imputation and the computation of the statistic one is interested in (cf. Eurostat 2002; Heeringa et al. 2010). In this section I shortly discuss each of these issues.

2.1 Sample design

The sample design can seriously affect the standard error. It is most likely that if a researchers assumes a simple random sample when the actual sample design involves clustering and stratification, standard errors will be wrong. Multi-stage designs involve several stages of sampling and sub-sampling and start from the random selection of clusters of elements (e.g. municipalities, census sections, dwelling blocks), i.e. primary sampling units (PSUs). If the design consists of several stages, the next step consists of drawing a subsample within each cluster. The advantage of (geographical) clustering is that interviewers can collect the interviews in a limited number of geographical areas, reducing the costs of the survey (e.g. Sturgis 2004). However, a major disadvantage is that clustering can seriously increase the standard error if the variance within clusters is small compared to the between-cluster variance with respect to the variable of interest. Stratification has the opposite effect. Stratification serves the purpose of increasing the representativeness of the sample and decreasing the risk that some parts in the population remain unrepresented. In order to do so, the population is divided into exclusive groups (strata). Subsequently an independent sample is drawn within each of these strata. Especially if the variance between strata is large with respect to the relevant variable, stratification contributes to decreasing the standard error. Usually, the effect of stratification is larger in the case of a clustered sample (cf. Lee and Forthofer 2006; Howes and Lanjouw 1998; Kish 1965). Of course, the effect of the sample design can differ from one variable to another: clusters or strata may differ strongly in the case of one variable and be rather heterogeneous in the case of another. A crucial point is that if the

ratio of selected clusters at the first stage to the total number of clusters in the population is small, other stages than the first add little to the standard error (for a mathematical elaboration, see Kish 1965; Cochran 1977). Therefore, the common practice is to approximately assess the sampling variance by estimating only the variability among the PSUs, since this is the dominating component of the total variance. In that case it is assumed that PSUs have been sampled with replacement and that no subsampling within the selected PSUs has been applied (this is the so-called 'ultimate cluster' approach, cf. Eurostat 2002; Heeringa et al. 2010). As a result, for a good approximation only accurate information on the first stage of the sample design is needed, considerably simplifying documentation and computation needs.

Several authors have compared confidence intervals for poverty figures under the assumption of simple random sampling with confidence intervals taking the sample design into account. For instance, Howes and Lanjouw (1998) found for Pakistan and Ghana standard errors that are at least 20 percent higher if the sample design is properly taken into account. Jolliffe et al. (2004) found even a larger impact with standard errors between 1.87 and 2.21 times as large in the case of Egypt. Rodger and Rodgers (1993) found standard errors being 1.5 to three times as large for a poverty index estimated on the basis of US data. Furthermore, Howes and Lanjouw (1998) emphasised that taking the sample design only partially into account can be as misleading as not taking it into account at all. However, the latter is not corroborated by findings of Biewen and Jenkins (2006). They found for German and UK data that once clustering at the household level is accounted for, estimates of standard errors are almost identical to estimates taking account of the entire sample design. This finding is important, as also in the case of EU-SILC and other surveys on income and living conditions (such as those included in the *Luxembourg Income Study*) clustering at the household level can easily be taken into account, even in the absence of accurate sample design variables. Commonly, measures of poverty and income are calculated at the household level, and assigned to all individuals within that household. This implies that all individuals within that household share the same value, and that the intra-cluster correlation for those variables is therefore by definition equal to 1. If the variance between households accounts for most of the total variance, the lack of accurate sample design variables is less of a problem. Although this assumption should in principle be tested for every statistic and survey separately, if we find similar results as Biewen and Jenkins (2006), it may provide some more trust in confidence intervals which only take account of clustering at the household level. Nevertheless, it should be stressed that this observation relates only to analyses at the individual level of variables defined at the household level (such as inequality in net disposable household income).

2.2 Imputation and weighting

Apart from the sample size and the sample design, standard errors are also influenced by other sources of random error. Among others, these include imputation and weighting. When there is item non-response, sometimes values are imputed (cf. Kalton 1983). If imputation is based on a random procedure, it adds another source of random error. The neglect of imputation generally leads to an under-estimation of the variance: imputation can increase random error, but also the denominator of the variance estimate (the number of respondents, n) can be overestimated if imputed values are treated as observations. As a result, the computation of standard errors should take this into account (Shao 1996; Shao and Chen 1998). However, the inclusion of this source of error is not easy. First, depending on the estimation technique and available software, information is needed on (1) the response / non-response status; (2) the imputation method used and information on the auxiliary variables; (3) information on the 'donor'[1]; (4) and information on the imputation classes (Eurostat 2002). Second, standard estimation procedures in many software packages do not include routines for taking account of imputation for estimating the standard error. Nonetheless, the impact of imputation can be large if there is considerable item non response. A non-response rate of 30% may lead to an under-estimation of the standard error by 10-50% (Kovar and Whitridge, 1995 as cited in Eurostat 2002). There are few studies on the impact of imputation procedures on standard errors in the case of

[1] In some imputation methods, a missing value is imputed by taking the non-missing value of another (otherwise similar) observation; the latter is called the "donor".

poverty indicators. One simulation study by Alfons et al. (2009) on the Austrian EU-SILC 2004 data finds that the additional uncertainty introduced by imputation is very limited in the case of the EU at-risk-of-poverty indicator, but somewhat larger for the average equivalent disposable household income. However, it is unclear to what extent these results can be generalised to other indicators or other EU-SILC countries.

Weighting is another potential source of random error. Weights assign more relative importance to some observations than to others in order to restore imbalances in the sample and avoid biased estimates. In general, weights are used to counteract three types of imbalance: unequal selection probabilities, unit non-response and (remaining) differences between the sample and known population data (Kalton 1983; Heeringa et al. 2010). In some cases imbalances in the sample occur on purpose, for instance some small strata may be over-represented to enable reliable estimates of these strata. For obtaining population estimates, respondents are given weights which are inversely proportional to the probability of being selected. Weights are also used to counterbalance unit non-response. In that case response propensity scores are estimated, i.e. the probability that a selected unit will cooperate in the survey. Consequently, the probability weight is multiplied by the inverse of the response propensity score. Finally, with post-stratification weights are adjusted such that the estimated distribution corresponds to known population totals, usually regarding demographic variables such as age and sex. In doing so, post-stratification may also increase precision by compensating coverage errors in the sample frame (cf. Groves et al. 2009). Even though weights are aimed at reducing potential bias of survey estimates, they may substantially increase standard errors. This is especially the case if the variance of the weights is large. Therefore, it is important to take weighting into account when computing standard errors. In EU-SILC for all countries weights have been developed to counteract variations in selection probabilities and unit non-response as well as to bring some demographic estimates in line with external population data (cf. Eurostat 2010c).

2.3 Complex poverty measures

The formulae for calculating standard errors do not only depend on the sample design, imputation and weighting, but also on the computed statistic. Standard errors are more straightforward to compute for some indicators than for others. For instance, the standard error of a proportion or the mean is well known and can easily be adapted to more complex sample designs (Kish 1965; Cochran 1977). Many poverty indicators consist of headcounts with a fixed (i.e. not random) poverty threshold. For instance, the Europe 2020 deprivation indicator estimates the proportion of the population which is deprived of at least 4 out of 9 items. In other words, in this case, the standard error is equal to that of a proportion. The same holds for the proportion of households with very low work intensity. As long as the poverty threshold is not estimated from the survey data itself, all poverty measures of the family of the well-known FGT-class of poverty measures have standard error formulae similar to those of a proportion or mean (cf. Foster et al. 1984; Kakwani 1993; Jolliffe and Semykina 1999).

However, this is not always the case. In the case of the at-risk-of-poverty indicator, the poverty threshold is estimated on the basis of the survey data: it is equal to 60 percent of the median equivalent household income in the (weighted) sample. Over the past 15 years, using linearisation, formulae and software for computing standard errors of many commonly used poverty indicators have been developed, including those which rely on poverty thresholds estimated from the sample. Several authors have derived formulae for standard errors in the case that the poverty line is estimated as a share of average or median income (e.g. Preston 1995). Some authors have combined this issue with additional considerations such as stochastic dominance over a range of poverty lines (Davidson and Duclos 2000), the (complex) sample design (Zheng 2001), the complex sample design and the influence of raking (the use of weights to balance the sample) (Berger and Skinner 2003) or the fact that household size should be considered a random variable as well (Thuysbaert 2008). Recently, macros for the linearisation of all Laeken poverty indicators have been published for the statistical software package SAS (Osier 2009). Importantly, linearisation relies on asymptotic assumptions, i.e. assumptions regarding a sufficiently large sample size. For population totals based on samples such as those in EU-SILC, with thousands of households included, there is no problem. However, one should be more careful when the method is applied to relatively small subsamples (cf. Osier 2009).

Many statistical software packages (e.g. SAS, SPSS, Stata) enable in a user-friendly way the computation of standard errors for proportions and means while taking account of the sample design and weighting. Usually, first the survey design variables must be indicated using a specific command (e.g. the *svyset* command in Stata or *CSPLAN* in SPSS). Thereafter specific commands must be used for estimating means and proportions while taking the survey settings into account (e.g. *svy: mean* in Stata, *CSDESCRIPTIVES* in SPSS or *PROC SURVEYFREQ* in SAS). However, ready-made procedures to compute standard errors of more complex poverty indicators and inequality measures are not included in a standard way. Estimation procedures which take account of the sample character of the poverty line, the complex sample design and weighting have been implemented in the freely available software package DAD (Duclos and Araar 2006; Araar and Duclos 2009)[2]. Among others, DAD accommodates inference for the FGT class of poverty measures in the case of a relative (estimated) poverty line, while taking the sample design into account. More recently, most of the modules of DAD have been implemented in the software package Stata under the name DASP (Araar and Duclos 2007)[3].

Not all poverty indicators can be estimated using DASP (e.g. the relative median at-risk-of-poverty gap). In that case one could turn to SAS and make use of the macro's published by Osier (2009). Another possibility is to apply the bootstrap method (a resampling method), in which case no formulae for computing the standard error and confidence intervals have to be derived. Shao and Chen (1998) studied the reliability of the bootstrap approach for quantiles and the low income proportion (of which the EU at-risk-of-poverty rate is a specific application) and found that even with complex sample designs the bootstrap performed reasonably well. Davidson and Flachaire (2007), compared in a more recent paper asymptotic and bootstrap inference in a Monte Carlo-type experiment. They found that in the case of a complex sample design bootstrapping inequality measures leads to inference that is not accurate even for very large samples, whereas bootstrapping poverty indices leads to satisfactory results[4].

3 The sample design of EU-SILC: what do we know and what is available?

The European Union Statistics on Income and Living Conditions (EU-SILC) is the EU reference source for micro data on income and living conditions. The dataset includes internationally and cross-temporary comparable variables for all EU Member States and some other countries. Many EU indicators designed to monitor poverty and social inclusion in the EU are based on EU-SILC (e.g. European Commission 2006; Marlier et al. 2007).

The reference population of EU-SILC consists of private households residing in the participating countries at the moment of selection. Currently 31 countries are involved in the EU-SILC process, namely all EU Member States plus the four non-EU members Iceland, Norway, Switzerland and Turkey. Nevertheless, the 2008 cross-sectional user database (UDB) contains information on only 27 countries (excluding France, Malta, Switzerland and Turkey). For the analysis presented in this article, EU-SILC 2007 data of France and Belgium have been included as well: France because 2008 data is missing completely and Belgium because in contrast to the 2007 data, in the 2008 data the sample design variables are missing. Considerable differences between participating countries exist in terms of sample design, sample frame and data source (e.g. Eurostat 2010a). In some countries, single stage designs are in use, whereas in other countries two- or three-stage designs are employed. In some countries (notably Hungary and France) two and three-stage designs are combined, depending on the region (stratum) and panel. Most countries apply stratification on at least one stage. Both sampling with equal probabilities and probabilities proportional to size are in use and in some cases systematic

[2] http://132.203.59.36/DAD

[3] http://132.203.59.36/DASP/index.html

[4] For non-smooth indicators such as many of the Laeken poverty indicators the jackknife is not recommended (e.g. Shao and Chen 1998; del Mar Rueda and Muñoz 2011). In the case of the at-risk-of-poverty indicator (FGT0 and FGT1) the resulting standard errors using the bootstrap are very close to those obtained on the basis of linearisation using the DASP module for Stata (figures available from the author).

sampling is applied (for an overview by country, see the annex in Goedemé 2010b). Additionally, it must be noted that EU-SILC has an important panel component, with a 4-year rotational panel design in the great majority of countries. Also with regard to the sample frame important differences exist, with potential problems of representativeness in the German and the Dutch case (Goedemé 2010b). Response rates vary substantially across countries ranging from 95 percent in Romania to 55 percent in Denmark (Eurostat 2010a). Last but not least in a number of countries many (income) variables in EU-SILC are based on (probably more reliable) register data rather than survey data. Recently, Lohmann (2011) has shown that this difference in data collection methods may substantially affect EU-SILC estimates. Although all these issues are very important for interpreting EU-SILC estimates, the remainder of this article will focus more narrowly on the consequences of the sample design and weighting for estimated standard errors. Verma et al. (2010) provide a more complete overview of both sampling and non-sampling error in EU-SILC.

Table 1: Number of persons, households and PSUs in the EU-SILC UDB, the Eurostat EU-SILC dataset and the sample design as reported by the national statistical offices

country code	country	persons	households	PSUs in UDB	PSUs in Eurostat data	Reported number of PSUs
AT	Austria	13,631	5,711	5,711	5,711	5,711
BE07	Belgium	15,493	6,348	243	243	275
BE08	Belgium	15,108	6,300	6,300	6,300	275
BG	Bulgaria	12,191	4,344	506	1,415	1,415
CY	Cyprus	10,025	3,355	3,355	3,355	3,355
CZ	Czech Republic	26,933	11,294	2,362	2,364	2,362
DE	Germany	28,904	13,312	13,312	13,312	No information
DK	Denmark	14,836	5,778	5,778	5,778	5,778
EE	Estonia	13,032	4,744	4,744	4,744	4,744
ES	Spain	35,970	13,014	1,994	1,994	2,000
FI	Finland	26,481	10,472	10,472	10,472	10,472
FR07	France	25,907	10,498	9,017	9,017	349
GR	Greece	16,869	6,504	1,064	1,064	1,056
HU	Hungary	22,363	8,818	4,875	5,245	4,184
IE	Ireland	12,551	5,247	1,723	1,723	1,747
IS	Iceland	8,644	2,887	2,887	2,887	2,887
IT	Italy	52,433	20,928	749	749	912
LT	Lithuania	12,150	4,823	4,823	4,823	4,823
LU	Luxembourg	10,147	3,779	3,779	3,779	3,779
LV	Latvia	13,120	5,166	912	912	930
NL	Netherlands	25,448	10,337	462	462	463
NO	Norway	14,216	5,553	5,553	5,553	5,553
PL	Poland	41,200	13,984	468	5,093	5,912
PT	Portugal	11,786	4,454	541	541	542
RO	Romania	19,131	7,805	779	779	780
SE	Sweden	18,825	7,452	7,452	7,452	7,452
SI	Slovenia	28,958	9,028	774	1,672	2,799
SK	Slovakia	16,546	5,450	5,450	5,450	5,450
UK	United Kingdom	21,043	8,936	1,014	1,014	1,065

Notes: The number of PSUs has been counted after the application of stratification by the region variable (DB040, UDB), respectively the proper stratification variable (DB050, Eurostat data) in countries where the PSU variable (DB060) is not unique across strata; however in some countries PSUs have been regrouped to avoid splitting of PSUs due to households moving from one region to another (see Goedemé 2010b).

Source: EU-SILC 2008 (BE, FR: 2007) UDB, the specific dataset prepared by Eurostat, National Intermediate EU-SILC 2008 Quality Reports and personal communication with Eurostat and national statistical offices.

From Table 1, it can be seen that the sample size varies between close to 9,000 persons (3,000 households) in the case of Iceland and more than 50,000 persons (21,000 households) in the case of Italy. In countries where the sample design consists of several stages, the number of PSUs is substantially lower, ranging from 275 in Belgium to nearly 6,000 in Poland. The number of explicit strata at the first stage varies from 1 (no stratification) to over 500 strata in Hungary (see Table 2). Additionally, in many countries with systematic sampling, implicit stratification has been applied

(meaning that data have been ordered according to several criteria). This is for instance the case for the UK and Norway.

The results presented in this article are based on two different versions of EU-SILC: the EU-SILC UDB as available to the research community (henceforth: 'UDB'), and a more complete dataset prepared by Eurostat (henceforth: 'Eurostat data'). As mentioned earlier, the data in the UDB referring to the sample design are incomplete: for almost one third of countries PSUs cannot be accurately identified and for all countries the original stratification variable is missing for confidentiality reasons. As a proxy to the stratification variable, one could use a variable which contains the region of residence at the moment of the interview (in contrast to the stratum at the moment of selection). However, in that case the number of strata is seriously under-estimated. In contrast to the EU-SILC UDB, the Eurostat data contain the proper stratification and PSU variables. Nevertheless, even in the case of the Eurostat data some discrepancies exist between the data and the reported number of primary strata and PSUs (see Tables 1 and 2). In addition, in the Eurostat data several strata contained only one PSU. By lack of information on the nature of these PSUs (self-representing or not[5]), with the exception of the UK data, all strata which contain one PSU have been re-grouped on the basis of (geographical) proximity and average equivalent net disposable household income [6] (cf. Eurostat 2002). As far as the UK EU-SILC is concerned, the single PSU is a self-representing PSU (i.e. Northern Ireland) in which a simple random sample of households has been drawn.

[5] If a PSU is self-representing, it is included with a probability equal to 1, which means that the PSU is rather a stratum than a PSU. For variance calculations this makes a difference with the case in which a stratum in the dataset contains only one PSU which has been selected (or contains respondents) among a larger amount of PSUs which populate the stratum in reality, but were not selected in the sample.

[6] For an in-depth discussion on the quality of the sample design variables in the 'EU-SILC Eurostat' dataset, see Goedemé (2010a). For a more precise description of the problems associated with the sample design variables in the UDB and the use of the region variable as a stratification variable, see Goedemé (2010b).

Table 2: Number of strata in the EU-SILC UDB, the Eurostat EU-SILC dataset and the sample design as reported by the national statistical offices

country	UDB	Eurostat data	Reported sample design
AT	3	247	247
BE07	3	11	11
BE08	3	11	11
BG	2	56	56
CY	1	9	9
CZ	8	53	53
DE	1	1	?
DK	1	1	1
EE	1	3	3
ES	18	93	93
FI	1	26	26
FR07	22	22	86
GR	4	90	90
HU	3	526	529
IE	1	138	138
IS	1	1	1
IT	5	288	288
LT	1	7	7
LU	1	160	160
LV	1	4	4
NL	1	40	40
NO	1	1	1
PL	6	211	211
PT	1	7	7
RO	8	88	88
SE	1	1	1
SI	1	6	6
SK	1	48	48
UK	1	31	31

UDB: number of strata as identified by the region variable (DB040)

Source: EU-SILC 2008 (BE, FR: 2007) UDB, the specific dataset prepared by Eurostat, National Intermediate EU-SILC 2008 Quality Reports and personal communication with Eurostat as well as national statistical offices.

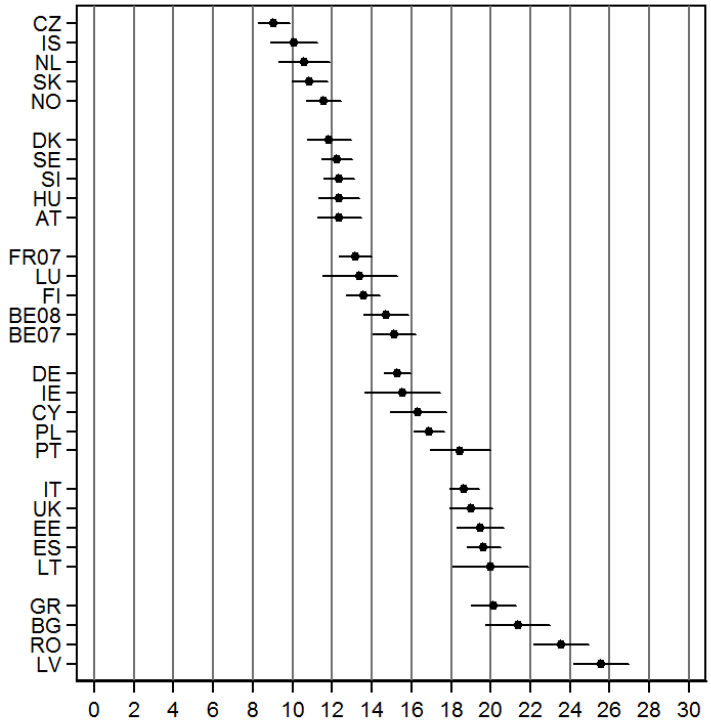
4 The Europe 2020 poverty reduction indicators

Sample design effects depend on the variable(s) one is interested in as well as on the computed statistic. As it is impossible to test the effect of various assumptions with regard to the sample design on the estimated standard error of every variable and statistic in EU-SILC, three indicators have been chosen which are related to the main fields of interest covered by EU-SILC: income, deprivation and labour market participation. More in particular the analysis is focused on the indicators which define the poverty reduction target as established by the European Council in June 2010 (European Council 2010): the at-risk-of-poverty rate, an indicator of severe material deprivation and an indicator of the

work intensity of the household. The target is defined as the absolute number of the population which is poor or socially excluded according to at least one of these three indicators. Nevertheless, in this article the three indicators are analysed separately.

Being at-risk-of-poverty means living in a household with an equivalised net disposable household income below 60 percent of the national median. Household income is equivalised using the modified OECD equivalence scale which attaches a weight of 1 to the first adult, a weight of 0.5 to all other household members aged 14 and over and a weight of 0.3 to household members aged less than 14. The equivalised household income is obtained by dividing total household income by the sum of the individual equivalence weights. All household members 'receive' the same equivalised household income. In other words, it is assumed that the living standard of all household members is the same (e.g. Atkinson et al. 2002; Marlier et al. 2007). In line with Eurostat practice, no top-bottom coding of income has been applied[7]. It must be stressed that in all countries, except Ireland and the United Kingdom, the income reference period is equal to the calendar year preceding the survey year. Severe material deprivation is measured by an index of nine items relating to financial stress and the enforced lack of some durables. All persons living in a household which at the moment of the interview lacks at least 4 out of 9 items are considered severely materially deprived. The list of items as well as the threshold is the same across all EU Member States (cf. Wolff 2010; Guio 2009; and for a theoretical discussion Goedemé and Rottiers 2011). The third indicator relates to the work intensity of the household. It is calculated by adding up the total number of months all household members at working age have worked during the income reference period, expressed in full-time equivalents. This is divided by the total number of months they could have worked. If the ratio is below 0.20, the household is considered to have a very low work intensity.

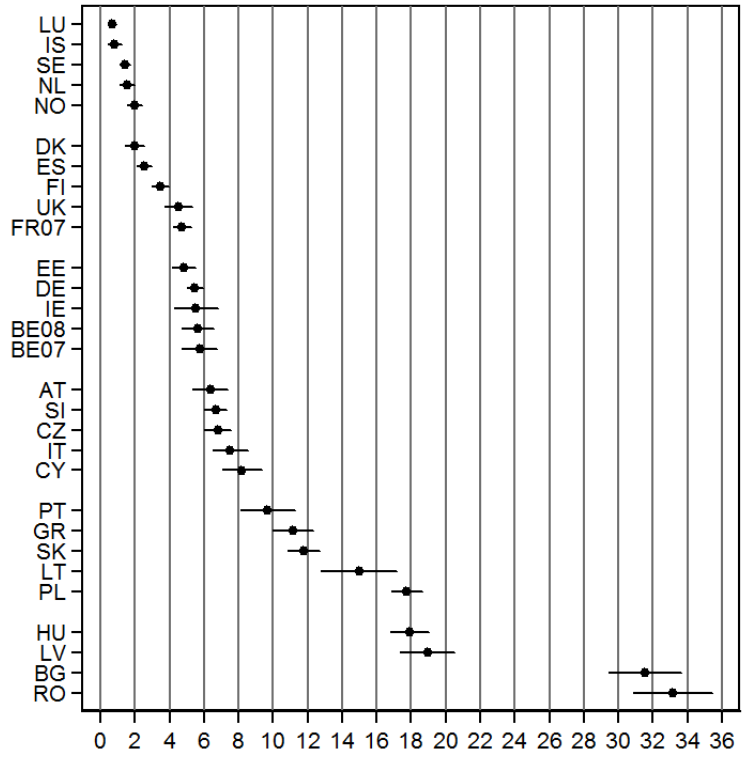
Fig. 1 At-risk-of-poverty rate with 95% confidence interval



Source: 'Eurostat EU-SILC' 2008 (BE, FR: 2007), own calculations.

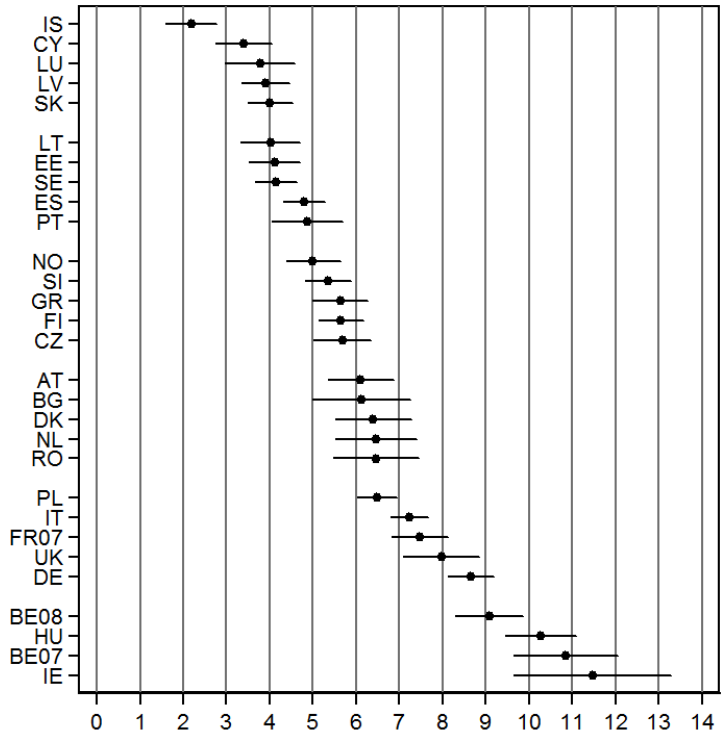
[7] In most countries many types of top-bottom coding would not make a big difference: neither for the estimated number of poor (cf. Van Kerm 2007), nor for the estimated standard errors (figures available from the author).

Fig. 2 Severe material deprivation rate with 95% confidence interval



Source: 'Eurostat EU-SILC' 2008 (BE, FR: 2007), own calculations.

Fig. 3 Share of population living in household with very low work intensity with 95% confidence interval



Source: 'Eurostat EU-SILC' 2008 (BE, FR: 2007), own calculations.

Figures 1-3 depict the point estimates for all three indicators together with 95% confidence intervals. The confidence intervals are estimated on the basis of the best available sample design information in the dataset prepared by Eurostat while using the DASP module for Stata (which means for nearly all countries full information on primary strata and almost complete information on primary sampling units, see column 'Eurostat data' in Tables 1 and 2 and the description of scenario 4 in section 5). These are the estimates which function as the benchmark for all other estimated standard errors presented in the next section. The at-risk-of-poverty rate ranges from 9 per cent in the Czech Republic to 26 per cent in Latvia. Compared to the at-risk-of-poverty indicator, the differences between European countries are much larger in the case of the indicator of severe material deprivation. In Luxembourg and Iceland less than 1 per cent of the population is estimated to be severely materially deprived compared to 33 per cent in Romania. The variation in estimates for the population living in a household with a very low work intensity is somewhat in between, ranging from less than 4 per cent in Iceland, Luxemburg, Cyprus and Latvia to over 10 per cent in Hungary, Belgium (2007) and Ireland. Overall, the country rankings differ much between the indicators. Rankings are most equal in the case of the at-risk-of-poverty and the severe material deprivation rate (Spearman's rank correlation coefficient of about 0.5). Also at the micro level income, deprivation and work intensity are only weakly correlated (absolute values of correlation coefficients vary between 0.10 and 0.52)[8].The precision of the estimates strongly depends on the indicator and the country under consideration. The width of the 95% confidence intervals ranges between 1.4 and 3.8 percentage points in the case of the at-risk-of-poverty rate, between 0.5 and 4.6 percentage points in the case of the deprivation rate and between 0.9 and 3.6 percentage points in the case of the low work intensity indicator.

[8] Estimates by country can be found in Goedemé (2010b). The weak correlation between deprivation and income poverty has been extensively documented in the literature, e.g. Dewilde (2004, 2008) and Whelan and Maître (2007).

5 The effect of making simplifying assumptions

In section two it has been argued that taking account of the sample design is of crucial importance for estimating standard errors accurately. However, as has been shown in section three, in many cases adequate information with regard to the sample design is lacking in the EU-SILC UDB. In this section I illustrate the importance of various assumptions with regard to the sample design. Additionally, the standard errors based on the EU-SILC UDB will be compared to those obtained using more complete information available in the 'Eurostat EU-SILC data'.

Standard errors have been estimated along four different scenarios. In scenario 1 (labelled 'persons'), it is assumed that EU-SILC consists of a simple random sample of individuals. In scenario 2 (labelled 'households') it is assumed that EU-SILC consists of a simple random sample of households. In scenario 3 (labelled 'UDB') the sample design variables of the EU-SILC UDB are taken into account, which means that whenever applicable – and as far as it is possible with the EU-SILC UDB – account is taken of stratification and clustering at a higher level than the household level. As discussed in section three, this information is very incomplete (see column 'UDB' in Tables 1 and 2). Finally, in scenario 4 (labelled 'Eurostat data') the more complete sample design information available in the Eurostat EU-SILC data has been used to take as much as possible account of clustering and stratification.

In all four scenarios it is assumed that PSUs are sampled with replacement. Imputation, as well as systematic sampling are ignored, but account is taken of weighting and the characteristics of each indicator. Even though weighting is taken into account, the separate effect of poststratification has not been accounted for. The neglect of systematic sampling and poststratification probably leads to a (slight) over-estimation of the standard errors, whereas the neglect of imputation could result in a more serious under-estimation of standard errors. Neither systematic sampling, nor imputation can be easily handled by lack of sufficient information in the dataset as well as the lack of user-friendly software. For every scenario standard errors have been estimated for all three Europe 2020 poverty indicators. In the case of the at-risk-of-poverty rate, the fact that the poverty line is estimated from the data has been taken into account. The estimation procedure is based on linearisation, using the DASP module and the generic commands for the robust estimation of proportions in Stata (Araar and Duclos 2007).

Apart from gaining some insight into the overall precision of EU-SILC estimates, the aim of the analysis is threefold. Although all three indicators are analysed at the individual level, they are measured at the household level. Therefore, at least clustering at the household level should be taken into account when computing standard errors: household members do not form independent units of observation, but are clustered within households (cf. Biewen and Jenkins 2006; Verma et al. 2010). In order to show the importance of this issue, standard errors assuming a sample of households (scenario 2) will be compared to assuming a simple random sample of individuals (scenario 1). Second, EU-SILC is not the only dataset which can be used for the measurement of income and living conditions and which lacks accurate sample design variables. Therefore, it is interesting to see whether the finding of Biewen and Jenkins (2006) for UK and German data can be replicated with EU-SILC data or whether, as emphasised by Howes and Lanjouw (1998), good estimates of the standard error are can only be achieved if the entire sample design is taken into account (scenario 2 and 4; cf. section 2.1). Third, by comparing scenarios 2, 3 and 4 it will be possible to evaluate whether making as much as possible use of the incomplete sample design information in the UDB leads to better estimates of standard errors than simply accounting for clustering within households.

When household variables are analysed at the individual level, researchers should take account of clustering at the household level. Results presented in tables 3 to 5 indeed show that if clustering at the household level is ignored, in the case of all three indicators standard errors are seriously underestimated. On average, the standard error which takes account of clustering at the household level is about 70 per cent larger than the standard error which assumes a simple random sample of individuals. Usually, the effect is strongest in the case of the at-risk-of-poverty indicator and weakest in the case of the work intensity indicator. Clustering at the household level has the strongest effect in

Romania (standard errors double) and the weakest effect in Denmark (standard errors are around 30 per cent larger).

Table 3: A comparison of estimated standard errors of the at-risk-of-poverty rate, EU-SILC 2008 (BE, FR: 2007)

		Standard error by scenario			
	point estimate	persons	households	UDB	Eurostat data
scenario		(1)	(2)	(3)	(4)
AT	12.36	0.33	0.57	0.57	0.57
BE07	15.13	0.30	0.53	0.56	0.55
BE08	14.72	0.32	0.58	0.57	n/a
BG	21.36	0.40	0.78	0.93	0.83
CY	16.34	0.40	0.73	0.73	0.73
CZ	9.06	0.22	0.39	0.42	0.41
DE	15.29	0.22	0.35	0.35	n/a
DK	11.84	0.42	0.57	0.57	n/a
EE	19.46	0.40	0.61	0.61	0.61
ES	19.65	0.24	0.43	0.44	0.44
FI	13.56	0.30	0.45	0.45	0.44
FR07	13.15	0.25	0.43	0.42	n/a
GR	20.14	0.32	0.59	0.59	0.59
HU	12.34	0.25	0.47	0.49	0.49
IE	15.53	0.48	0.96	0.98	0.98
IS	10.09	0.38	0.60	0.60	n/a
IT	18.67	0.20	0.36	0.39	0.39
LT	19.99	0.52	0.97	0.97	0.97
LU	13.40	0.53	0.95	0.95	0.95
LV	25.57	0.37	0.71	0.72	0.72
NL	10.59	0.35	0.54	0.66	0.66
NO	11.55	0.32	0.45	0.45	n/a
PL	16.88	0.20	0.39	0.40	0.40
PT	18.45	0.40	0.75	0.78	0.78
RO	23.57	0.32	0.69	0.74	0.72
SE	12.25	0.27	0.41	0.41	n/a
SI	12.33	0.23	0.38	0.37	0.40
SK	10.87	0.24	0.47	0.47	0.46
UK	19.00	0.29	0.54	0.55	0.56

n/a: 'Eurostat EU-SILC' dataset does not offer additional information. Standard errors based on linearisation using the DASP module for Stata. The fact that the poverty line is estimated from the data has been taken into account.

Source: EU-SILC 2008 (BE, FR: 2007); own calculations.

From Tables 3 to 5 it is also apparent that in many cases standard errors estimated while accounting for clustering within households (scenario 2) correspond relatively closely to standard errors taking as much as possible account of the sample design (scenario 4). For all three indicators, the estimated standard error in scenario 2 is on average less than 10 per cent below the standard error estimated in scenario 4. Even for many countries where at the first stage a sample has been drawn of large primary sampling units such as (groups of) municipalities or census areas, the estimated standard errors are relatively close to each other in both scenarios, repeating the finding of Biewen and Jenkins (2006). Nevertheless, this finding cannot be generalised: in several cases relatively large discrepancies can be found (e.g. the Netherlands with regard to the at-risk-of-poverty indicator; Italy, Portugal, Belgium 2007 and Romania in the case of the deprivation indicator; and the Netherlands and Belgium 2007 for the work intensity indicator). In other words: as stressed by Howes and Lanjouw (1998), whenever good sample design variables are available, they should be used for estimating standard errors and confidence intervals.

Table 4: A comparison of estimated standard errors of the severe material deprivation rate, EU-SILC 2008 (BE, FR: 2007)

		Standard error by scenario			
	point estimate	persons	households	UDB	Eurostat data
scenario		(1)	(2)	(3)	(4)
AT	6.36	0.27	0.52	0.52	0.52
BE07	5.74	0.21	0.34	0.52	0.52
BE08	5.64	0.23	0.48	0.47	n/a
BG	31.53	0.49	0.94	1.08	1.08
CY	8.20	0.32	0.58	0.58	0.58
CZ	6.81	0.19	0.34	0.39	0.39
DE	5.46	0.16	0.24	0.24	n/a
DK	1.97	0.21	0.28	0.28	n/a
EE	4.85	0.23	0.35	0.35	0.35
ES	2.55	0.11	0.20	0.23	0.22
FI	3.47	0.18	0.25	0.25	0.25
FR07	4.71	0.17	0.27	0.27	n/a
GR	11.17	0.30	0.54	0.61	0.60
HU	17.89	0.29	0.56	0.62	0.57
IE	5.53	0.32	0.59	0.65	0.64
IS	0.82	0.13	0.20	0.20	n/a
IT	7.53	0.16	0.30	0.55	0.52
LT	14.97	0.59	1.11	1.11	1.11
LU	0.68	0.08	0.13	0.13	0.13
LV	18.95	0.38	0.69	0.81	0.80
NL	1.55	0.14	0.23	0.23	0.23
NO	1.96	0.15	0.22	0.22	n/a
PL	17.75	0.22	0.42	0.44	0.46
PT	9.69	0.32	0.62	0.81	0.81
RO	33.16	0.42	0.90	1.22	1.17
SE	1.44	0.10	0.16	0.16	n/a
SI	6.67	0.19	0.32	0.33	0.33
SK	11.76	0.26	0.48	0.48	0.48
UK	4.50	0.19	0.39	0.41	0.41

n/a: Eurostat database does not offer additional information. Standard errors based on linearisation using Stata.

Source: EU-SILC 2008 (BE, FR: 2007); own calculations.

Does this also apply to incomplete and sometimes inaccurate sample design variables? The answer can be found when comparing scenarios 2 (households) 3 (UDB) and 4 (Eurostat data). In the case of four countries, the UDB contains full information on the sample design, as the sample design consists of a simple random sample of households (Denmark Iceland, Norway and Sweden), so standard errors can be computed accurately directly from the UDB. Also in the case of Belgium (2008), France (2007) and Germany no information can be added using 'Eurostat data', as both the UDB and the 'Eurostat data' do not contain the necessary information on the sample design.

In most countries, the difference between third and fourth scenario consists of applying a more detailed stratification. In seven countries stratification is applied to a sample of households and in eleven countries stratification is applied in a multi-stage design with larger primary sampling units (see Tables 1 and 2). The impact of adding stratification is relatively limited. As a result, for all three indicators, independent of the level of clustering, the difference between the UDB estimates (scenario three) and estimates based on the Eurostat data (scenario four) is trivial (less than 2 per cent difference). The difference is somewhat larger in Belgium (2007) and Finland (job intensity) as well as Romania (deprivation), but still below 10 per cent. In almost all countries, for all three indicators, the standard errors in scenario three are closer to those obtained in scenario 4 ('Eurostat data') than the standard errors in scenario two (households). This is especially the case for samples with relatively

large primary sampling units (e.g. Belgium 2007, the Netherlands, and Romania). Within countries, the exceptions usually relate to only one indicator.

In four countries (Bulgaria, Hungary, Poland and Slovenia) both the number of strata and the number of PSUs increase when going from scenario three (UDB) to scenario four (Eurostat data). For all four countries, the difference between both scenarios tends to be larger. In Bulgaria and Hungary standard errors in scenario three overestimate the standard error compared to those obtained using the 'Eurostat data'. Nevertheless, scenario three outperforms scenario two for the Bulgarian data[9]. In the case of Poland and Slovenia standard errors using the UDB tend to underestimate the standard errors in comparison with the standard errors based on the Eurostat data. Similarly, simulations on confidential France 2008 data indicate that standard errors for France 2007 data are substantially underestimated, especially in the case of the at-risk-of-poverty rate and the deprivation rate. Nevertheless, in all three countries scenario three (UDB) outperforms scenario two (households), except for the at-risk-of-poverty indicator in Slovenia (where the difference between both scenarios is small).

Table 5: A comparison of estimated standard errors of the share of people living in household with very low work intensity EU-SILC 2008 (BE, FR: 2007)

		Standard error by scenario			
	point estimate	persons	households	UDB	Eurostat data
scenario		(1)	(2)	(3)	(4)
AT	6.10	0.26	0.39	0.39	0.39
BE07	10.85	0.28	0.49	0.64	0.62
BE08	9.08	0.26	0.40	0.40	n/a
BG	6.12	0.25	0.49	0.61	0.58
CY	3.41	0.21	0.34	0.34	0.33
CZ	5.68	0.19	0.32	0.34	0.34
DE	8.66	0.19	0.27	0.27	n/a
DK	6.39	0.37	0.44	0.44	n/a
EE	4.11	0.22	0.31	0.31	0.31
ES	4.80	0.14	0.23	0.24	0.24
FI	5.64	0.21	0.28	0.28	0.26
FR07	7.48	0.21	0.33	0.33	n/a
GR	5.63	0.22	0.31	0.32	0.32
HU	10.26	0.23	0.41	0.46	0.42
IE	11.47	0.44	0.86	0.93	0.93
IS	2.19	0.21	0.30	0.30	n/a
IT	7.23	0.14	0.22	0.26	0.23
LT	4.02	0.23	0.35	0.35	0.35
LU	3.78	0.29	0.42	0.42	0.41
LV	3.90	0.18	0.27	0.28	0.28
NL	6.45	0.27	0.36	0.48	0.47
NO	5.00	0.23	0.32	0.32	n/a
PL	6.48	0.14	0.23	0.23	0.23
PT	4.87	0.23	0.39	0.41	0.41
RO	6.46	0.21	0.41	0.51	0.50
SE	4.14	0.17	0.24	0.24	n/a
SI	5.35	0.18	0.26	0.26	0.26
SK	4.01	0.16	0.27	0.27	0.27
UK	7.97	0.23	0.44	0.46	0.45

n/a: Eurostat database does not offer additional information. Standard errors based on linearisation using Stata.

Source: EU-SILC 2008 (BE, FR: 2007); own calculations.

[9] Please note that in the case of Hungary also the 'Eurostat data' do not contain fully accurate sample design variables.

6 Discussion and conclusion

If estimates are based on samples, they should be accompanied by appropriate standard errors and confidence intervals. This is true for scientific research in general, and is even more important if these estimates are used to inform and evaluate policy measures such as those aimed at attaining the Europe 2020 poverty reduction target. In order to compute accurate standard errors, the sample design, weighting, imputation and the complexity of the indicator should be taken into account. This requires proper sample design variables in the dataset, good documentation, and adequate software which takes these issues into account. In this article I have complemented the existing documentation on the sample design of EU-SILC and I have argued that user-friendly software exists for taking account of the sample design and weighting for estimating standard errors and confidence intervals. In addition, the effect of various assumptions with regard to the sample design on estimated standard errors has been evaluated for the three Europe 2020 indicators of poverty and social exclusion.

The main findings of the analysis can be summarised as follows. First of all, it should be stressed that in all circumstances one should take account of clustering within households when variables that are measured at the household level are analysed at the individual level, which is the case for most poverty indicators, and all EU2020 indicators. Otherwise, standard errors are severely underestimated, regardless of the sample design. Second, in line with Biewen and Jenkins (2006), the analysis shows that when standard errors take account of clustering at the household level, they are in many cases a good proxy for the standard errors which take account of the full sample design. This gives some hope for EU-SILC countries which completely lack sample design information (e.g. Germany, Belgium 2008) and also for the analysis of other surveys where sample design variables are missing (e.g. those included in the Luxembourg Income Study Database). Nevertheless, at the same time it is shown that in some cases standard errors are still largely under-estimated, which means that whenever good sample design variables are available, they should be used (cf. Howes and Lanjouw 1998). Third, the latter recommendation even holds when sample design variables are incomplete or not fully accurate. More precisely, the results presented in this article indicate that in the case of many EU-SILC countries, reconstructing the sample design variables from the available information in the UDB often leads to more accurate standard errors, even though the difference with assuming a simple random sample of households is not always very large. A detailed description of how to reconstruct the sample design variables is provided in Goedemé (2010b). A do-file with the necessary code for reconstructing the sample design variables with Stata can be downloaded from the internet[10]. I would like to stress that with standard software packages currently available, it is not more difficult to take account of the entire sample design than to simply take account of clustering at the household level, provided good sample design variables are available.

To what extent are these recommendations also applicable to other analyses (other variables, other types of analysis (e.g. regressions), or other data sources.)? The results presented in this paper show that the effect of clustering (and stratification) depends on the variable of interest, the sample design and the country of interest. However, although all three indicators analysed here are not strongly correlated, in most cases results consistently indicate that making as much as possible use of the available sample design information leads to the most accurate standard errors. Therefore, one can have some confidence in the results presented here. Nonetheless, it offers no guarantee. In fact, in principle the accuracy of estimates of standard errors should be evaluated for every analysis separately. In addition, the Eurostat data which act as a benchmark do not provide good sample design variables for Hungary and further research is needed with regard to self-representing PSUs, especially in the case of the Italian and French datasets. Finally, the analyses do not take account of systematic sampling, post-stratification and imputation. Especially the latter could strongly affect estimated standard errors.

More generally, some conclusions can be drawn for researchers, data providers and policy makers. First, it should be stressed that random error is an issue of concern, even for point estimates for the total population. Depending on the indicator, between 1 in 10 (indicator of severe material

[10] http://doiop.com/SvysetEU-SILC2008

deprivation) and 1 in 4 (indicator of very low work intensity) of the country-by-country comparisons is not statistically significant at the 95% confidence level, despite the relatively large sample size of EU-SILC in many participating countries. Of course, especially with regard to subgroup analyses, researchers will sometimes be disappointed about the lack of precision of their EU-SILC estimates. Not much can be done about this. However, from a scientific point of view it is more honest to recognise that on the basis of the data a hypothesis cannot be tested than to treat uncertain findings as if they were exact figures. Additionally, the estimates presented in this article show that the standard error heavily depends on the variable and country of interest. As a result, I would not recommend to adhere to general guidelines about the precision of a dataset and always compute statistical tests and report the precision of every estimate of interest (taking the findings summarised previously into account). In addition, even though confidence intervals give only some idea of the random error associated with estimates, they may also act as a general warning for the exact use of estimates based on samples and make readers more cautious about both random and non-random error.

Second, given the importance of good sample design variables, data providers should ground the decision with regard to the provision of these variables on a balanced scientific discussion about the confidentiality risks entailed by the disclosure of this information. If the necessary variables cannot be included in the UDB, other approaches should be considered. For instance, replicate weights could be provided alongside the EU-SILC UDB. If constructed appropriately, replicate weights enable researchers to bootstrap their estimates taking account of the sample design and weighting schemes (cf. Kolenikov 2010; Heeringa et al. 2010)[11].

Third, obviously random error means an important challenge to the definition and evaluation of poverty reduction targets. On the basis of EU-SILC data, the European Council agreed to reduce the number of Europeans affected by poverty or social exclusion with 20 million. In the case of the total EU estimate, EU-SILC will provide relatively precise figures for evaluating progress towards reaching this target. However, the target has to be translated into national policies with member states formulating their own targets and sub targets (Frazer et al. 2010). This is much more problematic. For instance, in the case of Belgium between 2,036,005 and 2,334,322 persons can be estimated to live in poverty or social exclusion with 95% confidence (EU-SILC UDB 2008, own calculations). In the context of the Europe 2020 Strategy, the Belgian federal and regional governments aim to reduce this number with 380,000 persons by 2020 (National Reform Programme of Belgium 2011). However, if the variance remains the same, the EU-SILC estimate should change in the coming years by at least 210,000 persons in order to be able to say with 95% confidence that the difference between the two point estimates is statistically significant from *zero*[12]. In other words, if EU-SILC is to function as an accurate policy-monitoring tool for reaching such precise targets, its *effective* sample size should be increased dramatically. Given the financial constraints many of Europe's national statistical institutions currently have to face, this is rather unlikely to materialise in the near future. A more realistic solution, which is technically more demanding, could lie in the use of simulation methods as well as administrative data to estimate changes in relation to the Europe 2020 poverty reduction targets.

[11] It should be noted that Eurostat is currently working on these issues and several projects are running to improve the sample design variables as well as to evaluate the feasibility of various approaches which should allow researchers to properly estimate standard errors for all EU-SILC countries.

[12] Figure applies to a two-tailed test. For a one-tailed test at least a difference of around 177,000 persons is needed. The estimation does not take account of the fact that the proportion is partially based on a random poverty line. Please note that for inter-temporal comparisons spanning a period of less than 5 years, one should take account of the covariance between the two years under evaluation as a result of the panel character of EU-SILC. The present estimate ignores this potential covariance. In addition, in many countries PSUs have been selected for the entire duration of EU-SILC, so also for differences between point estimates separated by more than 4 years, one should take account of the entire variance-covariance structure.

References

Alfons, A., Temple, M., & Filzmoser, P. (2009). On the Influence of Imputation Methods on Laeken Indicators: Simulations and Recommendations. *Conference of European Statisticians, Neuchatel, Switzerland, 5-7 October 2009* (pp. 9).

Araar, A., & Duclos, J.-Y. (2007). *DASP: Distributive Analysis Stata Package*: PEP, CIRPÉE and World Bank, Université Laval.

Araar, A., & Duclos, J.-Y. (2009). DAD: A Software for Poverty and Distributive Analysis. *Journal of Economic & Social Measurement, 34*(2/3), 175-189.

Atkinson, A. B., Cantillon, B., Marlier, E., & Nolan, B. (2002). *Social Indicators: the EU and Social Inclusion*. Oxford: Oxford University Press.

Berger, Y. G., & Skinner, C. J. (2003). Variance Estimation for a Low Income Proportion. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 52*(4), 457-468.

Biewen, M. (2002). Bootstrap Inference for Inequality, Mobility and Poverty Measurement. *Journal of Econometrics, 108*(2), 317-342.

Biewen, M., & Jenkins, S. P. (2006). Variance Estimation for Generalized Entropy and Atkinson Inequality Indices: the Complex Survey Data Case. *Oxford Bulletin of Economics and Statistics, 68*(3), 371-383.

Cochran, W. G. (1977). *Sampling Techniques*. New York: John Wiley & Sons.

Davidson, R., & Duclos, J.-Y. (2000). Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality. *Econometrica, 68*(6), 1435-1464.

Davidson, R., & Flachaire, E. (2007). Asymptotic and Bootstrap Inference for Inequality and Poverty Measures. *Journal of Econometrics, 141*(1), 141-166.

de Vos, K., & Zaidi, A. M. (1998). Poverty measurement in the European Union: Country-specific or Union-wide poverty lines? *Journal of Income Distribution, 8*(1), 77-92.

del Mar Rueda, M., & Muñoz, J. F. (2011). Estimation of poverty measures with auxiliary information in sample surveys. *Quality and Quantity, 45*(3), 687-700.

Dewilde, C. (2004). The Multidimensional Measurement of Poverty in Belgium and Britain: A Categorical Approach. *Social Indicators Research, 68*(3), 331-369.

Dewilde, C. (2008). Individual and institutional determinants of multidimensional poverty: A European comparison. *Social Indicators Research, 86*(2), 233-256.

Duclos, J.-Y., & Araar, A. (2006). *Poverty and Equity. Measurement, Policy, and Estimation with DAD*. New York: Springer.

Efron, B., & Tibshirani, R. J. (1998). *An Introduction to the Bootstrap*. Boca Raton: Chapman & Hall/CRC.

European Commission (2006). *Portfolio of Overarching Indicators and Streamlined Social Inclusion, Pensions, and Health Portfolios*. Brussels: European Commission.

European Council (2010). *European Council 17 June 2010 Conclusions*. Brussels: European Council.

Eurostat (2002). *Monographs of official statistics. Variance estimation methods in the European Union*. Luxembourg: Office for Official Publications of the European Communities.

Eurostat (2010a). *2008 Comparative EU Intermediate Quality Report. Version 2 - June 2010*. Luxembourg: Eurostat.

Eurostat (2010b). *Combating poverty and social exclusion. A statistical portrait of the European Union 2010*. Luxembourg: Publications Office of the European Union.

Eurostat (2010c). *Description of Target Variables: Cross-sectional and Longitudinal 2008 operation (Version January 2010)*. Brussels: European Commission.

Foster, J., Greer, J., & Thorbecke, E. (1984). A Class of Decomposable Poverty Measures. *Econometrica, 52*(3), 761-766.

Frazer, H., Marlier, E., Natali, D., Van Dam, R., & Vanhercke, B. (2010). Europe 2020: Towards a More Social EU? In E. Marlier, D. Natali, & R. Van Dam (Eds.), *Europe 2020. Towards a More Social EU?* (pp. 15-44). Brussels: P.I.E. Peter Lang.

Goedemé, T. (2010a). *The construction and use of sample design variables in EU-SILC. A user's perspective*. Report prepared for Eurostat, Antwerp: Herman Deleeck Centre for Social Policy, University of Antwerp.

Goedemé, T. (2010b). The standard error of estimates based on EU-SILC. An exploration through the Europe 2020 poverty indicators. *CSB Working Paper Series, WP 10/09* (pp. 36). Antwerp: Herman Deleeck Centre for Social Policy, University of Antwerp.

Goedemé, T., & Rottiers, S. (2011). Poverty in the Enlarged European Union. A Discussion about Definitions and Reference Groups. *Sociology Compass, 5*(1), 77-91, doi:10.1111/j.1751-9020.2010.00350.x.

Groves, R. M., Fowler, F. J. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology (Second edition)*. New Jersey: John Wiley & Sons.

Guio, A.-C. (2009). What Can Be Learned From Deprivation Indicators in Europe? *Paper Presented at the Indicator Subgroup of the Social Protection Committee, 10th February 2009* (pp. 33).

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied Survey Data Analysis*. Boca Raton: Chapman & Hall/CRC.

Howes, S., & Lanjouw, J. O. (1998). Does Sample Design Matter for Poverty Rate Comparisons? *Review of Income & Wealth, 44*(1), 99-109.

Jolliffe, D., Datt, G., & Sharma, M. (2004). Robust Poverty and Inequality Measurement in Egypt: Correcting for Spatial-price Variation and Sample Design Effects. *Review of Development Economics, 8*(4), 557-572.

Jolliffe, D., & Semykina, A. (1999). sg117 - Robust standard errors for the Foster–Greer–Thorbecke class of poverty indices. *Stata Technical Bulletin, STB-51*, 34-36.

Kakwani, N. (1993). Statistical Inference in the Measurement of Poverty. *The Review of Economics and Statistics, 75*(4), 632-639.

Kalton, G. (1983). *Introduction to Survey Sampling* (Vol. 35, Quantitative Applications in the Social Sciences). Beverly Hills: Sage Publications.

Kangas, O., & Ritakallio, V.-M. (2007). Relative to What? Cross National Pictures of European Poverty Measured by Regional, National and European Standards. *European Societies, 9*(2), 119-145.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Kolenikov, S. (2010). Resampling Variance Estimation for Complex Survey Data. *Stata Journal, 10*(2), 165-199.

Lee, E. S., & Forthofer, R. N. (2006). *Analyzing Complex Survey Data. Second Edition* (Vol. 71, Quantitative Applications in the Social Sciences). Thousand Oaks: Sage Publications.

Lohmann, H. (2011). Comparability of EU-SILC survey and register data: The relationship among employment, earnings and poverty. *Journal of European Social Policy, 21*(1), 37-54, doi:10.1177/0958928710385734.

Marlier, E., Atkinson, A. B., Cantillon, B., & Nolan, B. (2007). *The EU and Social Inclusion. Facing the challenges*. Bristol: The Policy Press.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A Nonparametric Approach to Statistical Inference* (Vol. 95, Quantitative Applications in the Social Sciences). Newbury Park: Sage Publications.

OECD (2008). *Growing Unequal? Income Distribution and Poverty in OECD Countries*. Paris: OECD.

Osier, G. (2009). Variance Estimation for Complex Indicators of Poverty and Inequality Using Linearization Techniques. *Survey Research Methods, 3*(3), 167-195.

Preston, I. (1995). Sampling Distributions of Relative Poverty Statistics. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 44*(1), 91-99.

Rodgers, J. R., & Rodgers, J. L. (1993). Chronic Poverty in the United States. *The Journal of Human Resources, 28*(1), 25-54.

Shao, J. (1996). Invited Discussion Paper Resampling Methods in Sample Surveys. *Statistics: A Journal of Theoretical and Applied Statistics, 27*(3), 203-237.

Shao, J., & Chen, Y. (1998). Bootstrapping Sample Quantiles Based on Complex Survey Data Under Hot Deck Imputation. *Statistica Sinica, 8*(4), 1071-1085.

Sturgis, P. (2004). Analysing Complex Survey Data: Clustering, Stratification and Weights. *Social Research Update, 43*, 1-6.

Thuysbaert, B. (2008). Inference for the Measurement of Poverty in the Presence of a Stochastic Weighting Variable. *Journal of Economic Inequality, 6*(1), 33-55.

Trede, M. (2002). Bootstrapping inequality measures under the null hypothesis: Is it worth the effort? *Journal of Economics, 77*(Supplement 1), 261-282.

Van Kerm, P. (2002). Inference on Inequality Measures: A Monte Carlo Experiment. *Journal of Economics, 77*(Supplement 1), 283-306.

Van Kerm, P. (2007). Extreme Incomes and the Estimation of Poverty and Inequality Indicators from EU-SILC. *IRISS Working Paper Series* (pp. 51). Luxembourg: CEPS-Instead.

Verma, V., Betti, G., & Gagliardi, F. (2010). An assessment of survey errors in EU-SILC. *Eurostat Methodologies and Working Papers* (pp. 70). Luxembourg: Eurostat.

Whelan, C., T., & Maître, B. (2007). Income, Deprivation and Economic Stress in the Enlarged European Union. *Social Indicators Research, 83*(2), 309-329.

Wolff, P. (2010). 17% of EU Citizens were at-risk-of-poverty in 2008. *Statistics in Focus* (pp. 8). Luxembourg: Eurostat.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. New York: Springer.

Zheng, B. (2001). Statistical Inference for Poverty Measures with Relative Poverty Lines. *Journal of Econometrics, 101*(2), 337-356.