DE GRUYTER OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 123–136, http://dx.doi.org/10.1515/JOS-2017-0007

Survey-Based Cross-Country Comparisons Where Countries Vary in Sample Design: Issues and Solutions

Olena Kaminska[1] *and Peter Lynn*[1]

In multi-national surveys, different countries usually implement different sample designs. The sample designs affect the variance of estimates of differences between countries. When making such estimates, analysts often fail to take sufficient account of sample design. This failure occurs sometimes because variables indicating stratification, clustering, or weighting are unavailable, partially available, or in a form that is unsuitable for cross-national analysis. In this article, we demonstrate how complex sample design should be taken into account when estimating differences between countries, and we provide practical guidance to analysts and to data producers on how to deal with partial or inappropriately-coded sample design indicator variables. Using EU-SILC as a case study, we evaluate the inverse misspecification effect (*imeff*) that results from ignoring clustering or stratification, or both in a between-country comparison where countries' sample designs differ. We present *imeff* for estimates of between-country differences in a number of demographic and economic variables for 19 European Union Member States. We assess the magnitude of *imeff* and the associated impact on standard error estimates. Our empirical findings illustrate that it is important for data producers to supply appropriate sample design indicators and for analysts to use them.

Key words: Cross-national studies; imeff; multiple frame design; complex sample estimation.

1. Introduction

There are many examples of multi-country surveys that are designed specifically for the purpose of cross-national comparisons (Lynn et al. 2006; Smith 2010), though the challenges that must be met in order to provide useful comparability are considerable (Kish 1994, 1999). In order to provide a basis for unbiased estimation of between-country differences, such surveys apply a standard definition of the target population (Heeringa and O'Muircheartaigh 2010) and select a probability sample from that population (Häder and Gabler 2003; Lynn et al. 2007). As well as enabling unbiased estimation, cross-national surveys sometimes also aim to standardize the precision of estimates within each

[1] Institute for Social and Economic Research, University of Essex, Colchester, Essex CO4 3SQ, United Kingdom of Great Britain and Northern Ireland. Email: olena@essex.ac.uk and plynn@essex.ac.uk

Acknowledgments: This research was carried out under the award Analysis of Life Chances in Europe (ALICE), funded by the UK Economic and Social Research Council. The ALICE Principal Investigators were Richard Berthoud and Maria Iacovou. EU-SILC data was supplied by Eurostat. We are grateful to Vijay Verma for advice, to Alexandra Skew and Francesco Figari for helping to document and interpret the data, and to Alita Nandi and Steve Pudney for advice on Stata routines. The empirical study of mis-specification effects (sections 3 and 4 of this article) was presented at the Joint Statistical Meetings in Vancouver, Canada, in August 2010 and at the World Statistical Congress in Dublin, Ireland, in August 2011. Additionally, our thoughts on preparing sample design indicators for cross-national analysis (section 2) were presented at the Comparative Survey Design and Implementation (CSDI) workshop in London, UK, in March 2011.

country (European Commission 2013). One way to achieve this is to select one specific important statistic and develop a sample in each country such that it leads to a defined precision for the estimate of that statistic (European Commission 2013). This defined precision has to be common across countries. For a multi-purpose survey, a more appropriate method is to set a common effective sample size (Lynn et al. 2007; Gabler et al. 2006). Effective sample size indicates how many cases a simple random sample would need in order to have the same precision as a particular (complex) sample design.

These requirements for a standard population definition, a probability sample, and a required precision leave scope for sample designs to vary across countries. Kish (1989, 41) mentions ". . . the selection methods and the sample designs of the surveys whose results are compared need not be at all similar. If they are based on good probability methods, the sampling method for each can be entirely distinct. Actually, for each sample we should utilize whatever selection method is most appropriate, feasible, and efficient . . .". If countries implement the most efficient and appropriate sample design, then differences in geography, population distribution, available sampling frames, and survey systems make it inevitable that countries will vary in whether and how they use stratified sampling, clustering, and unequal selection probabilities.

Differences in sample designs need not be a problem for estimation, but appropriate estimation requires the existence of appropriate indicators of components of the sample design. Specifically, indicators are needed of the strata used in a stratified sampling design, of the primary sampling units (PSUs) used in a multi-stage design, and of the design weights used in a design with variable selection probabilities. Furthermore, these indicators must be in a form that reflects the sample design when viewed as a single multi-national sample. If sample design indicators are either not available or not in an appropriate form, this can cause problems for analysis. Alternatively, the data producer could supply analysts with replicate weights (Dippo et al. 1984) that have been produced in a way that appropriately takes into account all features of the sample design. However, it can be argued that using replicate weights places a slightly higher burden on the analyst. Cross-national survey data sets often have one or more of the following problematic features:

- the indicator of sampling stratum is set to 'missing' for countries that don't implement stratification,
- the PSU indicator is left with missing values for countries where a single-stage design is implemented,
- the weight variable is set to 'missing' in countries where the sample is selected with equal selection probabilities, and
- for either the stratum or PSU indicator, the same range of values may have been (partially) used in different countries.

The consequences can be either that the analyst fails to notice the problematic features, leading to incorrect results, or that the analyst chooses to carry out analysis that ignores one or more components of the sample design (for example, clustering may be ignored if the PSU indicator has missing values), leading at least to biased estimates of standard errors.

This article has two aims. We first explain how missing information from countries that omitted a particular sample design feature can be 'filled', and how variables can be recoded if national data sets have been prepared without regard for the requirements for a cross-national data set. This should be useful for users who encounter these problems, but more importantly for data release organizations that, by following these steps, can make it easier for users to account for complex sample design. We then apply the method developed in the first section to create the best possible information on stratification, clustering, and weighting for a large cross-national survey data set. Estimates that use our filled and edited sampling information are compared with those that ignore one or more of the sample design indicators. Specifically, we study misspecification effects if all or part of a complex sample design is ignored in the situation where countries have different sample designs. We examine country comparisons of means and their standard errors for a number of demographic and economic variables.

While in this article we refer to comparisons between countries, the methodology presented has broader application. It applies to any situation where sample designs differ between domains, and these domains are either combined or compared in analysis. Such domains might include regions of a country or strata in a multi-stratum sample.

2. Preparing Sample Design Indicators for Cross-National Analysis

A cross-national sample can be viewed as a special case of a multiple-frame sample. Multiple-frame samples use more than one sampling frame to represent a population (Hartley 1962). Most literature on multiple frames discuss cases where one frame covers all units and another frame is cheap but covers only a subset, or where two frames overlap (Hartley 1962; Cochran 1965; Lohr, 2007; Lepkowski and Groves 1986). A cross-country survey represents a different situation, specifically where none of the frames overlap. According to Hartley (1962), a multiple-frame sample should meet the following requirements:

1) each unit in the population of interest should belong to at least one of the frames, and
2) for each sampled unit, it should be possible to record whether or not it belongs to the other frame(s).

In the cross-national survey context, these requirements are clearly met if we can assume the frames to be non-overlapping. Furthermore, cross-national surveys can be viewed as Hartley's case number 1, where all domain sizes are known (i.e., country totals). According to Hartley (Hartley 1962, 204), in this situation the frames (countries) should be treated as strata. He then notes "In case 1 the estimation problem is reduced to the standard methodology for stratified sampling." Thus each frame (country) should be viewed as a top-level explicit stratum, between which sample designs can vary.

2.1. Cross-National Stratum Indicator

For cross-national analysis, a single stratum indicator is required that reflects the complete multi-frame design. This indicator should reflect the sampling strata within each country, and treat countries as the top level strata (as samples were selected independently in each country). It is important that each stratum from the cross-national perspective should take a unique value, and therefore if one country supplies a stratum

indicator taking values of 1 to 5 and another country uses 1–7 to indicate strata, the values should be recoded (for example the second country's strata should be coded as 6–12). Any country that does not use stratified sampling should be treated as a single stratum. Thus, in countries with stratification the cross-national stratum indicator should take a different value for each national stratum, while for countries with no stratification, the cross-national stratum indicator should take the same value for each sample element. This is analogous to the situation in national surveys where some regions are treated as a single stratum, while others are subdivided into more detailed strata. If none of the countries has a stratified design, each country should be treated as a separate stratum and the stratum indicator for cross-national analysis should simply take a different value for each country.

2.2. *Cross-National PSU Indicator*

Analogously to the stratum indicator, the cross-national PSU indicator should indicate the units selected from each frame at the first stage of selection when the survey is viewed as a single cross-national sample. If none of the countries has a multi-stage sample design, the PSU indicator can be omitted with caution. Caution is needed in case there are multiple possible levels of analysis relating to hierarchically-associated units such as households and individuals. In this case, a single-stage sample of households would produce a multi-stage sample of individuals, where households are the PSUs within which individuals are clustered. In this situation, we suggest that the PSU indicator should be equivalent to a household indicator. Again, a different range of values should be used in each country so that each household has a unique value in the cross-national data set. Defined thus, the PSU indicator is important for individual-level analysis, while for analysis at household level it will, correctly, have no effect, as it will indicate the absence of clustering.

If all countries have a multi-stage design, then the cross-national PSU indicator should reflect this with a unique value for each PSU when the sample is viewed from a cross-national perspective. Attention is again needed to avoid the same value being used in more than one country.

In a situation where some, but not all countries use a multi-stage design, the indicator should take a unique value for each PSU in each multi-stage country, while it should take a unique value for each sample element in countries with a single-stage design. In this way, using the indicator will provide correct complex sample estimation in an analysis of multiple countries with and without multi-stage designs.

2.3. *Cross-National Weights*

For comparison of estimates between countries it is only necessary that the weight variable reflects the relative inclusion probabilities within each country; between-country differences in the mean weight will not affect comparisons for any type of ratio estimate such as means, proportions or model coefficients (Dorofeev and Grant, 2006, 82–84; see also Brewer 1963). However, we suggest routinely applying what we will call 'population scaling' to the weights. This will render them suitable for any kind of analysis, including that which combines countries, such as estimation for the total cross-national sample or

comparison of groups of countries. For unit i in country j the population-scaled weight for cross-national analysis should take the form:

$$w_{ij}^{s} = \frac{w_{ij}^{u} N_j}{\sum_{i=1}^{n_j} w_{ij}^{u}} \tag{1}$$

where w_{ij}^{u} is the national (unscaled) weight for the unit, and

n_j is the sample size in country j, and
N_j is the (assumed known) population size of country j.

Using this population-scaled weight, the weighted sample size for each country equals the population size of the country, that is $\sum_{i=1}^{n_j} w_{ij}^{s} = N_j$. An equivalent approach is used by the European Social Survey – see the description of "population size weight" in European Social Survey (2014).

In the special case where a country has a sample design with equal selection probabilities, the national weight may be missing. In this case, it should first be set to a constant value such as 1 for all sample elements in the country, that is $w_{ij}^{u} = 1 \,\forall\, i$. Then, Expression (1) can be applied though for such countries it can be simplified to:

$$w_{ij}^{s} = \frac{N_j}{n_j} \tag{2}$$

2.4. Cross-National Data Set

Once the steps outlined above have been followed, the three sample design indicator variables (stratum, PSU, and weight) are ready for use in any kind of cross-national analysis and can be incorporated into standard procedures for complex sample design estimation. Ideally, these steps should be carried out by the data production organization, so that data released to analysts is already in a suitable form for analysis. In that way, the analyst needs only to know how to carry out standard survey analysis, and does not additionally need to perform the data preparation relating to sample design.

3. Empirical Study of Misspecification Effects: Methods

Next, we study how important it is for an analyst of cross-national survey data to have full information on the complex sample design, and whether conclusions about differences between countries can be influenced by ignoring all or part of the sample design information. We concentrate on studying the effect of ignoring stratified and/or multi-stage (clustered) sampling where countries differ in their sample design, compared to estimation using stratum and PSU indicators that have been completed and edited following the procedures outlined in the previous section.

3.1. Data: EU-SILC

For our study we use data from the European Union Statistics on Income and Living Conditions (EU-SILC) survey. The EU-SILC has been carried out in all 27 EU Member States since 2007 (some started earlier) plus four non-Member States (Wolff et al. 2010). Both cross-sectional and longitudinal data are collected on income,

poverty, social exclusion, and other living conditions. Most items are collected through individual interviews with each adult in a household though some items are collected through a household interview. In most countries the data is collected by means of a survey with a rotating panel design (Iacovou and Lynn 2017). Though the details of the design vary, a typical design involves a four-wave rotation with annual interviews. Some countries select a sample of households via addresses, while others first select a sample of individuals and then identify the household of each selected individual. The latter group further subdivides into countries where all adult household members are interviewed and countries where only the selected individual is interviewed, as information on the other household members can be collected from population registers. Furthermore, some countries use a multi-stage clustered design, while others use a single-stage design. Key sample design parameters for each country are summarized in the supplemental data, Appendix 1 (available online at http://dx.doi.org/ 10.1515/jos-2017-0007).

We use data related to 2007, extracted from the longitudinal EU-SILC data set (EUSILC LONGITUDINAL UDB 2007 – version-1 of August 2009 [EOM]). The cross-sectional data set could not be used as it did not include a PSU indicator. We drop a number of countries from our analysis that either had not yet provided this data at the time of analysis, or for whom the indicators of sample design parameters – which are crucial to our analysis – were either missing completely or did not correspond to the description of the design (and where these discrepancies could not be resolved). This leaves 19 countries for analysis. The details can be found in the supplemental data, Appendix 1 (available online at http://dx.doi.org/10.1515/jos-2017-0007).

3.2. Data Editing: Complex Sample Design Variables

We apply the procedures outlined in Section 2 to the EU-SILC data. Although a majority of countries used stratified sampling, no stratum indicator exists in the data files, so we treat countries as strata and create a stratum indicator that takes a unique value for each country. For countries with single-stage sample designs we create a PSU indicator that is coterminous with household; for countries with multi-stage designs we use the existent PSU indicator, but recode to avoid between-country overlap in the ranges of values. We do not utilize weights provided by Eurostat, as these incorporate nonresponse adjustments for some countries, but not all, and do not always appear to reflect the described sample design. Instead, we derive our own design weights based on the documented description of the sample design in each country and, where relevant, the data item indicating the number of adults in the household. No attempt is made to develop nonresponse adjustments to these weights, as our focus in this article is on the effects of sample design on precision of estimates. For some countries, the weight for individual-level analysis is different from that for household-level analysis. Specifically, if a country implemented a sample of households, but only one individual was selected, we corrected for within-household selection for individual-level analysis (no such correction was needed for household-level analysis). If a country selected a sample of individuals and then included the household of each individual, we corrected for the fact that households are sampled with probability proportional to the size of the household (while no correction is needed for individual-level analysis). For further details please

see the supplemental data, Appendix 1 (available online at http://dx.doi.org/10.1515/jos-2017-0007).

3.3. Estimation

We use the svy commands in Stata 11.0 to provide estimates that take into account aspects of the sample design. Similar approaches can be used in other software packages. Our Stata syntax for estimating a difference between two countries in mean value of the variable var1 is as follows, where the variables strata1, psu, and weight1 are the three sample design indicator variables derived as described in the previous paragraph:

```
svyset psu [pw=weight1], strata(strata1)

svy: mean var1 if cntry1==1 | cntry1==2, over(cntry1)

lincom [var1]1 - [var1]2
```

It can be seen that this form of estimation is very simple to implement once the design variables have been correctly derived. We estimate differences between pairs of countries in a number of descriptive parameters (means and proportions, including some subgroup means). We note in passing that Stata estimation routines will incorrectly estimate the degrees of freedom used to construct the design-based confidence interval for the difference between countries whenever the true degrees of freedom differ between the countries. The effect is likely to be negligible when the degrees of freedom are large, but may not be negligible if the design in at least one of the countries has a small number of degrees of freedom. This problem exists independently of whether or not the design is correctly specified and is therefore not the focus of this article. The interested reader is referred to Valliant and Rust (2010) for discussion of this issue.

Our objective is to estimate what we call the inverse misspecification effect, *imeff,* in a range of scenarios. The misspecification effect, *meff* (Skinner 1989), is the ratio of the true variance of a sample statistic under the complex sample design to the estimated variance, when ignoring all or part of the sample design. The *imeff* (which equals 1/*meff*) is useful because it indicates the factor by which the variance of the estimate is under- or overestimated. If *imeff* is over 1 the variance is overestimated, but usually *imeff* is under 1, which means that the variance is underestimated by a factor of *imeff*.

In all cases, we assume that weights are correctly specified in the analysis. We consider three likely forms of misspecification when using the EU-SILC data:

- failing to take into account that samples are selected independently in each country (i.e., failing to treat countries as strata),
- failing to take into account that the sample is clustered (i.e., treating the sample as if it were a single-stage design), and
- only partially taking into account that the sample is clustered (suboptimal specification of clusters), specifically, recognizing that individuals are clustered within households, but not that households may be clustered within larger PSUs.

In combination, this leads to five possible types of misspecification (Table 1). For each type of misspecification, we estimate *imeff* for each of 90 pairs of countries, specifically all

Table 1. Design misspecification scenarios.

Five types of misspecification:	
Type 1	Ignore independence of samples and ignore clustering
Type 2	Ignore independence of samples
Type 3	Ignore clustering
Type 4	Ignore independence of samples and only partially consider clustering
Type 5	Only partially consider clustering

the pairs that consist of one country with a multi-stage (clustered) design and one with a single-stage design. (Of the 19 countries available for analysis, ten had multi-stage design and nine had single-stage design.) For household-level analysis, only misspecification Types 1, 2, and 3 are possible, as clustering of individuals within households is not relevant to household-level estimation. We estimate differences between countries for five household-level variables (listed in Table 2) and fifteen individual-level variables (listed in Tables 3, 4, and 5), leading to 8,100 estimates of *imeff*.

4. Results

As described above, we carry out analysis for five household-level estimates for each of three types of misspecification and for fifteen individual-level estimates for each of five types of misspecification. This is done for all 90 country pairs. Overall, we find that the *imeff* is, in general, considerable when the clustering is not specified, whereas the effect of ignoring the stratification is negligible for most estimates. Thus, results for Type 1 and Type 3 misspecification (see Table 1) are very similar, as are results for Types 4 and 5, while all 1,530 estimates of *imeff* for Type 2 are in the range 0.98–1.00. Therefore, we present here only the results from misspecification Type 1 and Type 4, as these capture all of the important findings.

4.1. Household-Level Questions

Starting with Type 1 results for each of the five household variables, in Table 2 we present the mean *imeff* (across the 90 country pairs). These are in the range 0.70–0.90. However, we also present the minimum and maximum estimated *imeff* for each variable, and this shows that in specific pairwise comparisons, *imeff* can be as low as 0.07. This means that the true variance could be 14 times the size of the estimated one if the design is misspecified in this way, and standard errors could be nearly four times the size of the estimated ones.

Table 2. Results for five household-level variables: misspecification Type 1 over 90 country-pairs.

	$\overline{y_1 - y_2}$	$\overline{imeff}$	*s.d. (imeff)*	*Min. (imeff)*	*Max. (imeff)*
Income	19160.32	0.80	0.25	0.07	1.00
Capacity to afford holidays	0.25	0.71	0.20	0.33	0.96
Capacity to afford meals	0.12	0.81	0.14	0.54	0.99
Ability to make ends meet	0.06	0.83	0.15	0.43	0.99
Number of household members	0.28	0.87	0.11	0.55	1.00

Table 3. Results for twelve individual-level variables: misspecification Type 1 over 90 country-pairs.

	$\overline{y_1 - y_2}$	$\overline{imeff}$	*s.d. (imeff)*	*Min. (imeff)*	*Max. (imeff)*
Gender	0.024	2.31	0.34	1.73	3.08
Age	2.06	0.64	0.09	0.39	0.78
Equivalized disposable income	11,737	0.38	0.14	0.03	0.63
Education (ISCED)	0.099	0.55	0.19	0.06	0.81
Economic activity	0.070	0.76	0.16	0.27	0.97
Employment	0.044	0.73	0.15	0.41	1.01
Education (males)	0.097	0.74	0.22	0.09	0.97
Economic activity (males)	0.066	0.99	0.14	0.57	1.22
Employment (males)	0.039	0.81	0.11	0.62	1.00
Education (females)	0.112	0.77	0.23	0.13	1.00
Economic activity (females)	0.075	0.94	0.18	0.37	1.14
Employment (females)	0.052	0.86	0.13	0.51	1.06

4.2. Individual-Level Questions Available for All Household Members

Table 3 summarizes results for those individual-level estimates that are based on observations of all individuals in each sample household, either because all individuals were interviewed or because only one person was interviewed, but information for other individuals was obtained from a population register. Twelve of the 15 individual-level estimates are of this type, of which six are whole-sample, three are based on males only and three on females only. Among these estimates, the largest mean *meff* (across the 90 country pairs) is 2.31 for gender which is, unusually, (well) above the value of 1.00. This is a unique situation, which suggests that failing to take into account clustering results in an overestimate of the standard error of the difference. This reflects that PSUs (which, for several countries, consist of households) in the population are more heterogeneous with respect to gender than random samples of the same size from the whole population would be. As a consequence, sample-clustering reduces the standard error of the estimated gender distribution.

Apart from gender, the mean *imeff* (across the 90 country pairs) ranges from 0.38 for mean equivalized disposable income to 0.99 for the proportion of males who are economically active. This is a much greater range than observed above for household-level estimates, reflecting the larger intra-cluster correlation for individual variables due to the additional level of clustering (individuals within households) and the larger sample size per PSU. Failing to correctly take clustering into account is therefore particularly problematic for individual-level estimation. Some values of *imeff* for differences between two countries are very low indeed, with the smallest being 0.03 for a difference in mean equivalized disposable income, implying that standard errors could be underestimated by a factor of six. As an indicator of the extent to which this underestimation may affect analytical conclusions, we would note that, excluding gender, 27 of the 990 comparisons (2.7%) appear significant ($P < 0.05$) if the design is misspecified in this way, but not significant if correctly specified.

Unlike Type 1 misspecification (Table 2), which completely ignores clustering, Type 4 misspecification partially accounts for clustering. Specifically, household IDs are treated as clusters in both countries and this is compared to correctly specifying PSUs in countries

Table 4. Results for twelve individual-level variables: misspecification Type 4 over 90 country-pairs.

	$\overline{y_1 - y_2}$	$\overline{imeff}$	*s.d. (imeff)*	*Min. (imeff)*	*Max. (imeff)*
Gender	0.024	0.95	0.06	0.78	1.02
Age	2.06	0.93	0.12	0.59	1.07
Equivalized disposable income	11,737	0.77	0.26	0.07	1.00
Education (ISCED)	0.099	0.72	0.24	0.08	0.97
Economic activity	0.070	0.90	0.19	0.33	1.10
Employment	0.044	0.82	0.16	0.45	1.03
Education (males)	0.097	0.78	0.23	0.10	0.97
Economic activity (males)	0.066	0.94	0.13	0.54	1.13
Employment (males)	0.039	0.86	0.10	0.65	1.00
Education (females)	0.112	0.79	0.23	0.14	1.00
Economic activity (females)	0.075	0.90	0.17	0.35	1.00
Employment (females)	0.052	0.88	0.13	0.53	1.06

where such are present. The same variables are used and the same comparisons are implemented as in Table 3.

As expected, the estimate of the difference itself is not influenced (Table 4). Overall, the mean *imeff* for Type 4 misspecification is much less pronounced than the mean *imeff* for Type 1 misspecification. It comes closer to 1.0 for all estimates except for two (economic activity for males and females), which were close to 1.0 already in Table 3 (the change for these two estimates is minor). For example, the mean *imeff* changes from 0.38 to 0.77 for equalized disposable income. The minimum and maximum *imeff* are also much closer to 1.0. Overall, taking into account clustering of individuals within households improves the estimates considerably, even when ignoring prior stages in a multi-stage sampling design.

4.3. *Individual-Level Questions Available for All Household Members in Some Countries and for One Household Member in Other Countries*

Thus far, we have discussed the situation in which information is available for all household members, obtained either through an interview or from a register. However, in countries where only one person was interviewed in each household, some variables were not available from a register, leading to a situation in which some variables (for example health evaluation) are only available for one household member. When using such variables to construct estimates of differences between countries, the effect of misspecification can be different from that of variables available for all household members, even though correct specification takes the same form. When comparing two countries, one with a multi-stage sample of households and one with a single-stage sample of households, we distinguish between four situations:

a) both countries may have one individual observed per household,
b) both have all individuals observed per household,
c) only the clustered country has all observed, or
d) only the unclustered country has all observed.

These four scenarios have potentially different implications for misspecification, so in Table 5 we present results separately for each scenario.

Table 5. Results for self-assessed general health (individual-level): misspecification Type 1.

		$\overline{y_1 - y_2}$	$\overline{imeff}$	*s.d. (imeff)*	*Min. (imeff)*	*Max. (imeff)*
All individuals in both countries (48 comparisons)	All	0.071	0.72	0.06	0.61	0.85
	Men	0.060	0.91	0.07	0.69	1.06
	Women	0.080	0.89	0.06	0.77	0.98
One per household in both countries (six comparisons)	All	0.057	0.95	0.01	0.93	0.97
	Men	0.050	0.98	0.01	0.97	0.99
	Women	0.063	0.97	0.03	0.94	1.00
All individuals in PSU country; one per household in non-PSU country (24 comparisons)	All	0.087	0.83	0.08	0.68	0.97
	Men	0.076	0.93	0.07	0.74	1.05
	Women	0.095	0.92	0.06	0.81	0.99
One per household in PSU country; all individuals in non-PSU country (twelve comparisons)	All	0.071	0.83	0.03	0.77	0.89
	Men	0.063	0.96	0.02	0.93	0.98
	Women	0.079	0.95	0.03	0.92	1.00

It can be seen that values of *imeff* are modest when both countries interview only one person per household, but a little more substantial when one of the countries interviews all persons. The largest values of *imeff* arise when both countries interview all persons, as in this case an entire level of clustering is being ignored in both countries.

5. Conclusions

Our findings show that misspecification effects in cross-national comparisons can be considerable and can result in serious bias in standard errors of estimates of between-country differences. This would result in biased hypothesis testing (Type 1 errors). Bias is greatest when multi-stage sample selection is ignored completely in estimation. Bias is smaller, but still substantial (for individual-level estimates) when the first stage is ignored and only the clustering of individuals within households is acknowledged. Furthermore, misspecification effects have been shown to depend on the nature of the difference in sample design between the two countries being compared. The corollary of this is that in multi-country comparisons, if designs are misspecified in estimation, the chances of a country being identified as an outlier depends on the sample design adopted in that country. This is clearly undesirable.

To avoid misspecification effects in cross-national comparisons, it is necessary not only for sample design indicators (PSU, stratum, and design weight) to be present on the data set, but also for these indicators to be in a form that is suitable for cross-national analysis. Indicators that are suitable for national analysis of each country do not necessarily meet this requirement, but in Section 2 above we have set out the steps necessary to convert these indicators into a suitable form. These steps are not particularly demanding and we propose that they should be carried out by a relevant central agency before data is released to analysts. This is efficient, as it avoids duplication of effort, and mistakes by analysts who may not be experts in sample design. Once suitable indicators for cross-national analysis have been produced, correct specification can easily be achieved with standard software, leading to unbiased estimation of standard errors.

However, we are aware that the EU-SILC is certainly not the only cross-national survey data set in which the sample design indicators are not in suitable form. An analyst of any such data would be well-advised to follow the data preparation steps that we propose here. Furthermore, there are some cross-national survey data sets that do not release indicators of sampling strata or primary sampling units to secondary analysts at all. The European Social Survey is one prominent example (see http://www.europeansocialsurvey.org/data/). The producers of such data sets should be encouraged to release these indicators so that analysts can appropriately estimate standard errors and test hypotheses.

While we have focused here on how best to estimate the impact of sampling error on cross-country comparisons, the impact of other components of statistical error may be equally important. It is not our intention to imply otherwise. In addition, estimating the magnitude of error *post-hoc* is no substitute for controlling the error at the design and data collection stages. All sources of error (coverage, sampling nonresponse, measurement, editing, and so forth) should be given due attention within a total survey error framework (Biemer 2010; Groves and Lyberg 2010) that recognizes interactions and dependencies between the error sources. Our comments on sampling error should be considered within that context, although further discussion of the broader context is outside the scope of this article.

Finally, we should note some limitations of our research. We have not examined all possible variants of misspecification. In particular, we have not assessed the effects of ignoring variation in design weights. Nor have we assessed the effects of ignoring stratified sampling within countries. The first of these is, in general, likely to lead to even greater underestimation of standard errors. The second is likely to have a rather more modest effect in the opposite direction. Furthermore, we have examined a limited number of estimates for one survey, albeit important ones. Effects might be different in magnitude for estimates of substantially different parameters and for substantially different sample designs (e.g., those with much larger, or smaller, cluster sample sizes). However, we do not feel that any of these limitations invalidate our main conclusion, which is that misspecification can have a serious effect and can (and should) be avoided. Though the effect may be different in magnitude in other circumstances, the data preparation steps outlined here guarantee that the effects can be completely avoided. As implementing the steps has very modest resource implications, we think that this should always be done.

6. References

Biemer, P.P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74: 817–848. Doi: http://dx.doi.org/10.1093/poq/nfq058.

Brewer, K.R.W. 1963. "Ratio Estimation and Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process." *Australian Journal of Statistics* 5: 93–105.

Cochran, R.S. 1965. "Theory and Application of Multiple Frame Surveys." *Retrospective Theses and Dissertations*. Paper No. 4080.

Dippo, C.S., R.E. Fay, and D.H. Morgenstein. 1984. "Computing Variances from Complex Samples with Replicate Weights." In Proceedings of the Survey Research Methods Section of the American Statistical Association. 489–494.

Dorofeev, S. and Grant, P. 2006. *Statistics for Real-Life Sample Surveys: Non-Simple-Random Samples and Weighted Data*. Cambridge: Cambridge University Press.

European Commission. 2013. *Handbook on Precision Requirements and Variance Estimation for ESS Household Surveys*. Luxemburg: Publications Office of the European Union. Available at: http://ec.europa.eu/eurostat/documents/3859598/5927001/KS-RA-13-029-EN.PDF (accessed January 2017).

European Social Survey. 2014. *Weighting European Social Survey Data*. Available at: www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data_1.pdf (accessed 28 October 2015).

Gabler, S., S. Häder, and P. Lynn. 2006. "Design Effects for Multiple Design Samples." *Survey Methodology* 32: 115–120.

Groves, R.M. and L. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74: 849–879. Doi: http://dx.doi.org/10.1093/poq/nfq065.

Häder, S. and S. Gabler. 2003. "Sampling and Estimation." In *Cross-Cultural Survey Methods*, edited by J.A. Harkness, F.J.R. Van de Vijver, and P.Ph. Mohler, 117–134. Hoboken, New Jersey: Wiley.

Hartley, H.O. 1962. "*Multiple Frame Surveys*." In Proceedings of Social Science Section of American Statistical Association meetings. Minneapolis, Minnesota. Available

at: http://ww2.amstat.org/sections/srms/Proceedings/y1962/Multiple Frame Surveys.pdf (accessed January 2017).

Heeringa, S.G. and C. O'Muircheartaigh. 2010. "Sampling Designs for Cross-Cultural and Cross-National Survey Programs." In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, and T. Smith, 251–268. New Jersey: Wiley.

Iacovou, M. and P. Lynn. 2017. *Design and Implementation Issues to Improve the Research Value of the Longitudinal Component of EU-SILC*. Monitoring Social Inclusion in Europe, edited by A.B. Atkinson, A.-C. Guio and E. Marlier. Chapter 27. EU Publications.

Kish, L. 1989. Q/A 21.1 Comparisons of Surveys. *Questions/Answers. From the Survey Statistician*, edited by A.M. Vespa-Leyder, 40–41.

Kish, L. 1994. "Multipopulation Survey Designs." *International Statistical Review* 62: 167–186.

Kish, L. 1999. "Cumulating/Combining Population Surveys." *Survey Methodology* 25: 129–138.

Lepkowski, J. and R.M. Groves. 1986. "A Mean Squared Error Model for Dual Frame, Mixed Mode Survey Design." *Journal of the American Statistical Association* 81: 930–937.

Lohr, S. 2007. *Recent Developments in Multiple Frame Surveys*. In Proceedings of Survey Research Methods Section of American Statistical Association meetings, Salt Lake City, Utah. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y2007/Files/JSM2007-000580.pdf (accessed January 2017).

Lynn, P., L. Japec, and L. Lyberg. 2006. "What's So Special about Cross-National Surveys?" *Conducting Cross-National and Cross-Cultural Surveys: Papers from the 2005 Meeting of the International Workshop on Comparative Survey Design and Implementation (CSDI)*, edited by J. Harkness. ZUMA, Mannheim.

Lynn, P., S. Häder, S. Gabler, and S. Laaksonen. 2007. "Methods for Achieving Equivalence of Samples in Cross-National Surveys: the European Social Survey Experience." *Journal of Official Statistics* 23: 107–124.

Skinner, C.J. 1989. "Introduction to Part A." In *Analysis of Complex Surveys*, edited by C.J. Skinner, D. Holt, and T.M.F. Smith, 23–58. Chichester: Wiley.

Smith, T.W. 2010. "The Globalization of Survey Research." In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P.Ph. Mohler, B.-E. Pennell and T.W. Smith, 477–484. Hoboken, New Jersey: Wiley.

Valliant, R. and K.F. Rust. 2010. "Degrees of Freedom Approximations and Rules of Thumb." *Journal of Official Statistics* 26: 585–602.

Wolff, P., F. Montaigne, and G.R. González. 2010. "Investing in Statistics: EU-SILC." In *Income and Living Conditions in Europe*, edited by A.B. Atkinson and E. Marlier, 37–55. Luxembourg: Publications Office of the European Union.

Received December 2015
Revised June 2016
Accepted June 2016