Distr. GENERAL
02 September 2013
WP 25
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Seminar on Statistical Data Collection
(Geneva, Switzerland, 25-27 September 2013)

Topic (iv): Multiple modes of data collection

ANALYSIS OF NON-RESPONSE IN THE ESS – STATE OF PLAY, STATE OF THE ART AND SOME GUIDELINES FOR IMPROVEMENT

Working Paper

Prepared by Martin Karlberg and Pilar Rey Del Castillo - Eurostat

I. Introduction

1. Non-response is a perennial problem for official statistics, and the general perception is that the non-response rate is increasing over time.

2. To take stock of the situation, Eurostat has commissioned a major study which provides an extensive overview of the state of play concerning non-response across the European Statistical System, as well as the state of the art concerning methods for reducing, treating and reporting non-response. To allow for comparability across Member States, the study focuses on four surveys: the HBS, EU-SILC, LFS and SBS. This paper provides a summary overview of the findings of the study.

II. Qualitative overview of the ESS state of play regarding non-response

3. Based on (i) a questionnaire to ESS NSIs and (ii) available methodological reports, the study provides an overview of methods to reduce non-response, treatment of unit and item non-response, and reporting of non-response.

A. Methods to reduce non-response

4. One widely practiced method is the sending of **introductory or reminder letter** to the participants is a widely extended practice. It is carried out in almost all countries for EU-SILC and SBS, about two thirds of the countries for LFS and in slightly more than one half of the countries for HBS.

5. More than two thirds of the countries report **making several contacts** with the respondents, either through visits, telephone calls or e-mail contacts. Among the countries reporting several contacts, the usual practice is to make three or more attempts (up to six attempts in one country for EU-SILC).

6. On the contrary, **interviewer training** is reported by practically every country, some of them mentioning the existence of an Interviewer Handbook. (In one country, incentives is provided to interviewers for completed interviews in LFS and EU-SILC.)

7. On a more political level, participation is **compulsory** in some countries. For the LFS, this holds for about for half of the countries, whereas this is nearly never the case for SBS and HBS.

8. In HBS, it is common to **substitute** for non-respondents in the sample originally selected due to the high rates of non-response. Typically, these substitutions are made after matching with characteristics of the non-respondents.

9. **Incentives** to respondents are rarely used (five countries do it for EU-SILC and 6 do it for HBS), and only a handful of countries practice **translation** of the questionnaires.

B. Treatment of unit non-response

10. The vast majority of countries apply some unit non-response adjustment method to most of the four surveys.

1. Treatment of unit non-response

11. **Weighting** and **calibration** are the most popular techniques. For SBS, calibration is rarely applied; instead, various **imputation** methods are applied (often based on data from administrative sources, and/or previous years).

2. Treatment of partial unit non-response

12. For partial unit non-response (individuals in a household not responding), three different approaches are applied for EU-SILC: (i) **Full-case imputation**; (ii) **Adjustment factor** applied to the total income based on the characteristics of the household and the non-interviewed person; (iii) **exclusion of all data** for households with responses missing for one or more persons.

3. Treatment of item non-response

13. The approach for treating item non-response varies across the four surveys and – across countries. Sometimes, imputation is applied, sometimes not. In contrast to most surveys, the situation for SBS is less problematic, as imputation is already applied in the case of unit non-response; when a variable is missing for an enterprise, its value is in general completed using administrative registers.

C. Reporting of non-response

14. For EU-SILC, the non-response indicators are governed by a Commission Regulation 28/2004; the three indicators are based on combinations of (i) the address contact rate R_a, (ii) the household response rate R_a and (iii) the individual response rate R_p.

15. For LFS, Commission Regulation (EC) No 377/2008 provides one global non-response rate, which is subdivided by reason into (i) refusals, (ii) non-contacts and (iii) other reasons. With few exceptions, this is presented at household level (some countries present non-response rates at the individual level). The treatment of non-response in the **follow-up waves** is also different between countries. Some participating countries do not take previous non-response into account when calculating the non-response in later waves, whereas others do.

16. For SBS, Commission Regulation (EC) No. 275/2010 establishes that Member States shall report the **weighted unit non-response information** on the activities covered by the NACE codes in scope. (Special provisions apply if if several surveys/administrative sources are used for certain characteristics.)

17. Due to the substitution frequently applied in HBS, the computation of non-response rate is complicated; in many situations the extent and procedures for substitution are not sufficiently controlled, and adequate records of sample implementation not maintained. Sometimes, different response rates are reported for different parts of the survey (e.g. the diary).

III. Quantitative analysis of the non-response rate in the ESS

18. In the study, a rich dataset, with non-response rates from all surveys over 5-10 years, is analysed.

D. Overall non-response level and taxonomy

19. SBS surveys have, in general, low rates of response, whereas HBS has, by far, the highest rate of non-response among the four surveys. Within surveys:

    a. The level of LFS non-response is considerably higher in countries where participation is voluntary.

    b. In the panel survey EU-SILC, new samples have a higher non-response rate than the whole sample (the observed difference exceeds 5 percentage units). The principal cause of non-response in new samples is the non-response in contacted houses.

    c. For LFS, which reports non-response by reason, the main reason for non-response is refusals for countries with voluntary participation and non-contacts in countries with mandatory participation.

E. Trend over time

20. For LFS and HBS, the fitted models indicate an increase in non-response over time in the order of one percentage unit per year (½ percentage units for LFS in those countries where this survey is mandatory). In contrast, no clear time trend is discerned for EU-SILC and SBS.

F. Mode effects

21. Concerning the effect of the data collection mode, some vague indications of a mode effect are observed in the LFS (non-response in the order of 10 percentage units higher for CAPI compared to CATI and PAPI). However, the quality of the data gathered for this exercise is not sufficient for any firm conclusions to be drawn. For EU-SILC and SBS, no mode effect was seen (similar data quality issues prevent firm conclusions here as well, though). For HBS, mode data were only available for roughly half of the countries, so no mode effect analysis was made.

IV. State of the art regarding methods for addressing non-response

22. The study provides a review and inventory of existing documentation on non-response. Non-response being a hot topic in survey methodology for a long time, there is a large body literature, and a complete listing of documents was deemed infeasible. A selection has been done following the two-fold criteria of the relevance of the contents and the publishing date, with a preference for the more recent publications. Sections G through J provide an overview of inventory conducted.

G. Basic issues

23. Non-response occurs when elements of the population selected in the sample do not provide the requested information. If the population element does not provide any information at all, unit non response occurs, while when some information is provided but some questions in the survey remain unanswered, there is a case of item non-response. This is a key distinction because there are different methods developed specifically for each type of nonresponse.

24. Non-response affects not only the precision of the estimates but can make them biased when the (unobserved) values of the non-respondents systematically differ from those of the respondents for the survey variables. Because non-response errors usually account for a substantial part of non-sampling errors the response rate is one of the main quality indicators for a survey. A high non-

response rate may lead to a large bias, which, contrary to the standard error of the estimator, does not depend on the sample size.

25. To successfully prevent and treat nonresponse it is necessary to understand the various types of mechanisms by which missing data can emerge. From this point of view, three types of nonresponse are distinguished:

    - Missing Completely At Random (MCAR): when the missingness of a variable data is not related to its unknown value and to the values of the other variables in the same unit, for example, when the respondent or the interviewer accidentally overlooks a question. In this case the missing values are a random sample of all the values and the results of the analysis of the observed data will not be biased. The problems that could arise are mainly an issue of reduced precision.
    - Missing At Random (MAR): when the nonresponse is related to the observed data in the same unit but not to the unknown value of the missing variable. A typical example is an elderly respondent with difficulties to remember the answer to a question. In this case the missingness is a random process conditional on the observed data –conditional on the age in the example– being the missing values a random sample of the missing variable within the classes defined by the other observed variables.
    - Missing Not At Random (MNAR): when the nonresponse is related to the unknown value of the variable itself. For example, a respondent can feel his real answer as socially not acceptable and eludes responding. No simple solution exists for this case and serious bias may occur.

26. Most of the methods developed to treat the nonresponse problem tackle the MAR missingness (and also the MCAR, being a special case of MAR). The MNAR missingness requires sophisticated techniques such as developing specific models that describe the relation between the variables or, alternatively, conducting sensitivity analysis to assess the performance of different assumptions about the missing data mechanism.

27. Special cases of item nonresponse have been studied in the last decades. For example, in panel or longitudinal surveys part of the initial sample not respond after a certain point in time and this is called *panel mortality* or *panel attrition*. Specific adjustments for this problem have been studied and developed.

28. Another type of item nonresponse that is acquiring increasing interest with the web surveys is the *breakoff*: the respondent starts to respond to the survey but fails to complete it resulting in missing data for the last items of the survey. An issue currently under discussion is whether separate treatment to adjust for this breakoff must be applied, or whether this could be done in the normal framework. To resolve this issue, the causal mechanism resulting in breakoff must be studied.

H. Preventing non-response

29. The literature studies the non-response problem from two different approaches: the prevention before it occurs and the treatments required accounting for it once nonetheless has occurred. If nonresponse could be completely prevented the problem would disappear, so considerable resources are currently devoted to improving data collection procedures in order to reduce it. The methods for prevention rely on a body of knowledge from within the behavioural sciences and entail a number of techniques to be applied before or during the data collection process.

30. The main sources of non-response are non-contacts and refusals. Non-contacts are generally attributable to physical impediments to reach the respondent or to the call-pattern established by the interviewer. A third source of non-contacts is also been considered in an increasing number of countries: the language barrier.

31. Refusals are a somehow more complicated problem to deal with. The reasons that cause persons or businesses to not respond are usually of a behavioural kind, and very often related to sociological or psychological factors. Several theories have been developed to explain this issue, often attributing different levels of influence to the role of the interviewer. But all of them accept that the experience

of the interviewer as well as his positive attitude in relation to the survey can help to create a good interaction between the interviewer and the respondent, thereby increasing the response rate.

32. It is also generally admitted that the mode of data collection plays an important role in survey participation, as do other characteristics of the survey, such as the length of the interview. Literature discussing the pros and cons of the different measures that could be taken to reduce non-response is very extensive. The number of publications dealing with the advantages and disadvantages of the different modes of data collection, including mixed modes, is also significant. Moreover, information about the non-respondents can come up useful for the reduction of bias.

I. Treatment of non-response

1. Treatment of unit non-response

33. The bias generally caused by non-respondent units has to be adjusted. To this end adjustment techniques based on the use of auxiliary information are used. This information consists of variables external to the survey, related to the target variables or to the response behaviour, and whose distribution over the population or across the complete sample is known. The auxiliary variables can be available for all the population units, for all sample units, or at aggregated population level. Auxiliary variables available at sample level are often referred to as *paradata*.

34. A widely used technique for the treatment of unit non-response is post-stratification. The population is divided into strata defined by combinations of a set of auxiliary variables, and weights are corrected within each stratum. The implementation of this technique requires a certain minimum number of sample units within each stratum, and complete information on the population regarding the frequency for each stratum.

35. A more general approach based in the generalized regression estimation and not requiring complete auxiliary information for the entire population distribution is the *linear weighting* technique. It consists of estimating a linear model predicting the target variable of the survey from a set of auxiliary variables. The correction weights are calculated as the sum of the corresponding weight coefficients. The generalized regression estimator is an asymptotically design unbiased estimator of the population mean of the target variable, which makes this technique particularly useful for large samples. Post-stratification can be considered as a particular case of linear weighting.

36. *Multiplicative weighting*, or *raking*, is an alternative technique that can be used only when all the auxiliary variables are qualitative. Both linear and multiplicative weighting appear as particular cases of the *calibration technique*, an iterative process based on Lagrange method and producing unbiased estimators.

37. The input to adjustment weights is ultimately provided by the auxiliary variables. The more informative these auxiliary variables are, the better the adjustment. Variables strongly related both to the target variables of the survey and to the response behaviour are good candidates to be used in reducing the non-response bias. The selection of the appropriate auxiliary variables is thus a core issue in non-response treatment. Criteria such as the minimising of the maximal absolute bias or the variance of calibrating weights are frequently used in assessing the effectiveness of auxiliary variables for the reduction of the response bias, with the purpose to arrive at a good set of selected variables. For the determination of the disaggregation level with which they should be introduced, it is usual to employ ad-hoc techniques or methods based on classification trees, such as CHAID or CART.

38. An alternative method for the treatment of non-response is the *response propensity* technique. It is frequently used to solve the problems of undercoverage caused by self-selection in internet panels. The *propensity score* is defined as the conditional probability that a sample unit responds to the survey, given the corresponding values of a set of variables or attributes, which –just because this method is mainly applied to web surveys- are sometimes called *webographic* or *psychographic* variables. It is assumed that sample units with similar propensity scores are also similar with respect

to the attributes on which the propensity score has been measured. Propensity scores can be estimated with *logit* or *probit* models and they can be used as weights in the estimation of the target variables. The resulting technique has some similarities with raking, or multiplicative weighting, with the advantage that it can also be used for continuous variables.

2. Treatment of item non-response

39. Item non-response can happen for different reasons: a particular question may be unanswered because it is a sensitive one, because the respondent does not understand the question or does not know the answer, or simply because this question has been accidentally skipped. But missing data can also occur as a consequence of data editing if the collecting agency decides to treat wrong answers as unknown.

40. Due to missing values the analysis of the data may be hindered. The option used by most of the statistical software packages –to delete records containing missing data or at least those containing missing values in the variables under analysis– is not the best of solutions, as it can significantly reduce the information available and introduce a bias. For this reason, it is a regular practice to impute the missing data on the basis of auxiliary variables that have been observed. The result of imputation is a complete data set in which missing values have been replaced by imputations. One of the advantages of this strategy is that it allows for splitting up nonresponse adjustment and data analysis into two different stages. Datasets including imputed data may be analysed using various standard techniques and possible inconsistencies stemming from heterogeneous treatment of missing data would not occur.

41. *Model-based methods* in the wide sense include all the imputation methods making assumptions about the probability distribution of the variables and/or about the missingness mechanism. They constitute the state of the art in missing data imputation for the statistical inference community. A milestone in this field is the *Expectation Maximization (EM) algorithm*. It provides an iterative procedure to compute the likelihood of a sample that has missing data.

42. Currently, *Multiple Imputation (MI)* methods appear to be the most frequently used model-based procedures for handling missing data in multivariate analysis. They consist of first replacing each missing value by a set of imputations drawn from an assumed model distribution and thereafter combining these sets in a particular way. One of its claimed advantages is that only a small number of imputations (between three and five) are needed in order to obtain relatively efficient estimators. Many MI methods have been developed using different models assumptions for continuous, categorical and mixed continuous and categorical data. Statistical software packages often now include MI as one of the options and this is making a big contribution to its dissemination. Nonetheless, there is controversy about its actual efficiency and the unverifiable assumptions on the models. Another important point to emphasize is that in many applications the issue of non-response bias is much more crucial than that of variance thus rendering the main (efficiency) argument supporting MI less important.

43. Another more advanced approach to imputation is the use of classification and prediction techniques from the machine learning field: neural networks, genetic algorithms, Bayesian networks, classification trees, fuzzy sets, hybrid procedures etc. Generally speaking, prediction methods are used for the imputation of continuous variables, while classification methods are applied to categorical or otherwise discrete variables.

44. Machine learning procedures seem easier to automate, what makes them especially appropriate to be used for massive imputation tasks. This could be a solution to problems such as integration of data from different surveys, small area estimation and others that must be addressed in the domain of statistical surveys.

J. Evaluation of non-response

45. Indicators of the non-response error may be useful when comparing different surveys sharing the same target population or different waves of a longitudinal survey. They can also help monitoring the data collection process.

46. The *Non-response rate* has been widely used as such an indicator. It is simple, easy to interpret and its calculation is straightforward. Its main drawback is that it just measures the potential effect of nonresponse bias and is thus not identical to the actual nonresponse error. An alternative could be to use as indicator the *non-response bias* itself but this bias is rarely directly measurable. Moreover, when the number of survey estimates is high the multivariate information provided would be difficult to interpret, and some synthetic measure would have to be constructed.

47. For all these reasons a number of alternatives to the non-response rate have recently been proposed. Among them the representativeness indicators, or *R-indicators*, based on the concept of *representative response*, seem to be the most popular. It refers to the degree of similarity between the set of respondents and the complete sample, measured in relation with selected characteristics of the population described by a set of socio-demographic or socioeconomic auxiliary variables. To construct R-indicators, the *response propensity* is defined for each sample unit and each auxiliary variable as the conditional distribution of the individual response indicator on the variable.

48. The *R-indicator* is defined as one minus two times the standard deviation of the response propensity, always taking values between 0 and 1. Some interesting properties of this indicator have been proven. Its drawback derives from the fact that R-indicators depend strongly on the auxiliary variables that are chosen, which can be a particularly significant limitation for comparisons between different surveys. While these indicators are being employed as a useful complement to the non-response rate, as they provide valuable insight on the dissimilarities between the respondents and the selected sample concerning the selected auxiliary variables, they are not meant to be used as tool for non-response adjustment.

V. Guidelines for improvement of the current non-response practice

49. The guidelines are based on a comparison of the state of play in the ESS (Sections II and III) to the state of the art in non-response research (Section IV).

K. Preventing non-response

50. It is suggested that modern (e.g. Internet collection and use of administrative sources) are employed more frequently – typically in a “mixed mode” setting. The quality of the mode data gathered does not allow the quantitative analysis (Sections III) to draw any major conclusions concerning the mode effect. (While not explicitly brought up in the study, the appropriate and authoritative evaluation of mode effect involves embedded experiments; any effect estimates in “observational study” data will have a bias of unknown magnitude and direction.)

51. It is also suggested that the number of contact attempts is increased before a sample unit is considered to be a non-respondent. (While not explicitly brought up in the study, this could be combined by sub-sampling prior to the final contact attempt(s); this might complicate estimation, but could enhance representativity.)

52. From a purely quantitative perspective, the quantitative analysis demonstrated that declaration of the participation in a survey as mandatory (even when no fine or other penalty is applied) has a positive impact on the response rate. However, other considerations obviously come into play, making any such recommendation quite sensitive.

L. Treatment of non-response

53. While **calibration** is commonly applied for the treatment of unit non-response in social surveys, the actual application varies considerably across countries and surveys. As calibration variables and calibration techniques are selected to optimise calibration at a country level, the study argues that the lack of harmonisation may have an impact on the comparability among countries, and proposes that for the sake of comparability, Member States should try to converge on this matter by applying more similar calibration methodologies on the same set of calibration variables.

54. For SBS, the treatment of unit non-response seems less harmonised across countries than that in the social surveys. Some countries treat it by carrying forward the value from the previous year (if available), whereas other countries complete the missing information by means of administrative registers. The application of re-weighting also seems to vary across countries. If possible (considering the heterogeneous situation w.r.t. the access to administrative data), unit non-response treatment in the SBS should be more harmonised.

55. While it already proposes (section K) that a shift towards **administrative data** is made to minimise non-response and reduce the response burden, the study also proposes that administrative data are used to treat unit and item non-response once it has taken place to complete the missing information. **Linking** between different surveys (and with registers) is also encouraged (a prerequisite being a unique identifier, which is recommended – but the sensitivity of the issue is acknowledged). Statistical matching is also mentioned as a possible method, but its suggested that this be used in specific cases.

56. Due to the complexity of the issue of **imputation** and the heterogeneity of current practices – conditioned by the availability of complementary information to support the imputation procedures, the study suggests that this issue is addressed in working groups where the Member States would be able to discuss the imputation methods they are applying, as well as their advantages and disadvantages.

M. Evaluation of non-response

57. The quality of the information on non-response disseminated for each of the four surveys analysed in the study varies considerably. EU-SILC has the highest degree of harmonisation, whereas HBS is in the other end of the spectrum. To cope with this situation, the study suggests developing a standard methodological report template, including harmonised definitions of non-response rates, to be completed and disseminated by the countries. The methodology applied in the EU-SILC could be used to set a benchmark for this process.

58. To augment the non-response rates, more sophisticated indicators, such as the R-indicators (section J) could be considered. These alternative indicators could help complete the information provided by non-response rates and measure the consequences and knock-on effects of non-response on the sampling bias and sample quality. However, given the drawbacks associated with R-indicators, this could only be done for sets of surveys where it would make sense to select the same set of auxiliary variables.

VI. Concluding remarks

59. A major advantage of the present study is that it presents non-response issues in parallel for a set of surveys across a large number of ESS countries, and that it contains both a qualitative and quantitative overview of the situation in the ESS over the past 5-10 years, as well as an overview of the state of the art, which serves as the basis for recommendations for improvements.